Statistical Decisions

Wojciech Niemiro

December 31, 2020

Contents

Chapter 1

M-functionals and M-estimators

1.1 Loss functions and M-functionals

The general framework is the following. We assume that a real valued ,,loss"

$$\rho(a, Z)$$

depends on our decision a and on a (sampled value of) random variable Z. Let $\mathcal{A}$ be the set of available decisions (actions) and $\mathcal{Z}$ (a quite abstract) space of values of Z. Thus $\rho : \mathcal{A} \times \mathcal{Z} \to \mathbb{R}$ is a *loss function*. We are to minimize the expected loss denoted by

$$R(a) = \mathbb{E}\rho(a, Z).$$

If $\rho(a, Z) \geq 0$ then $R(a) \in [0, +\infty]$ is well defined (the expectation exists but can be infinite). Sometimes R is called a risk function. Assume that there exists $a_* \in \mathcal{A}$ such that

$$R(a_*) \leq R(a) \quad \text{for all } a \in \mathcal{A}$$

and $R(a_*) < \infty$. This is true in most examples we are to consider. Usually, we additionally assume that a_* is the unique minimizer, that is $R(a_*) < R(a)$ for $a \neq a_*$.

Of course a_* depends on the probability distribution of Z, because the expected value is computed with respect to this distribution. We say that a_* is an *M-functional* (Minimization-functional). This terminology is used even if the mininimizer is not necessarily unique, Below we show that some popular "characteristics of a probability distribution" are in fact M-functionals.

1.1.1 EXAMPLE (Expectation). Let $\mathcal{Z} = \mathcal{A} = \mathbb{R}$ and

$$\rho(a, z) = (z - a)^2.$$

Then $a_* = \mathbb{E}Z$ is the unique minimizer of $R(a) = \mathbb{E}(Z - a)^2$. Note also that the minimum value is $R(a_*) = \mathrm{Var}Z$. $\triangle$

1.1.2 EXAMPLE (Quantile). Let $\mathcal{Z} = \mathcal{A} = \mathbb{R}$ and $p \in]0,1[$ be fixed.

$$\rho(a,z) = a + \frac{z-a}{1-p}\mathbb{1}(z > a).$$

Then a_* is a minimizer of $R(a)$ iff $\mathbb{P}(Z < a_*) \le p \le \mathbb{P}(Z \le a_*)$. Thus, a_* is a p-th quantile of Z. In general, it is not unique. If we additionally assume that the c.d.f. $F(a) = \mathbb{P}(Z \le a)$ is continuous, then we simply have $F(a_*) = p$. Moreover, the minimum value also has a nice interpretation, namely $R(a_*) = \mathbb{E}(Z|Z > a_*)$. In the world of finance, the quantile a_* is known as VaR (Value at Risk) and $\mathbb{E}(Z|Z > a_*)$ as CVaR (Conditional Value at Risk). △

1.1.3 REMARK. If we consider an affine transformation of a loss function, that is define $\tilde{\rho}(a,z) = c + b\rho(a,z)$ with $b > 0$ then, of course, risk functions R and $\tilde{R}$ corresponding to ρ and $\tilde{\rho}$ have the same minimizer. For example, the loss function

$$\tilde{\rho}(a,z) = (a-z)(1-p) + (z-a)\mathbb{1}(z > a) = \begin{cases} (a-z)(1-p) & \text{for } z \le a; \\ (z-a)p & \text{for } z > a, \end{cases} \tag{1.1.4}$$

can be used in Example 1.1.2 to get p-quantile. Note that for $p = 1/2$ we obtain $\tilde{\rho}(a,z) = |a-z|/2$.

More generally, we can consider transformed loss of the form $\tilde{\rho}(a,z) = c(z) + b\rho(a,z)$ and still obtain the same minimizers of R and $\tilde{R}$ (that is, the intercept of the affine transformation can depend on z) This fact is useful to relax assumptions on the moments. For example, if $\rho(a,z) = |a-z|$ then we need to assume that $\mathbb{E}|Z| < \infty$ to define $R(a) = \mathbb{E}|a - Z|$. The minimizer of R is the median of Z, provided that Z has finite expectation. If we change ρ to $\tilde{\rho}(a,z) = |a-z| - |z|$, the minimizer of $\tilde{R}(a) = \mathbb{E}(|a-Z| - |Z|)$ is the median, without any restrictions on the moment.

1.1.5 EXAMPLE (Linear Regression, Least Squares). Let $Z = (X,Y)$, where X and Y are two one-dimensional random variables. The problem is to predict Y given the observed value of X. Suppose we consider linear predictors of the form $\hat{Y} = \beta_0 + \beta_1 X$. Then $a = (\beta_0, \beta_1)$ and the space of actions is $\mathcal{A} = \mathbb{R}^2$. The standard loss function is quadratic:

$$\rho(\beta_0, \beta_1, x, y) = (y - \beta_0 - \beta_1 x)^2.$$

We have to assume that $\mathbb{E}X^2, \mathbb{E}Y^2 < \infty$. The best linear predictor obtains for $\beta_1^* = \mathrm{Cov}(X,Y)/\mathrm{Var}(X)$, $\beta_0^* = \mathbb{E}Y - \beta_1^* \mathbb{E}X$. △

1.1.6 EXAMPLE (Linear Regression, Multivariate). Similarly as in the previous example consider $Z = (X,Y)$. Now assume that $X = (X_1, \dots X_d)^\top$ is a random vector and Y is one-domensional random variable. Linear predictors of Y given X can be written as $\hat{Y} = \beta_0 + \sum_{j=1}^d \beta_j X_j = \beta_0 + X^\top \beta$, where $\beta = (\beta_1, \dots, \beta_d)^\top$. For the quadratic loss, $\rho(\beta_0, \beta, x, y) = (y - \beta_0 - x^\top \beta)^2$, the best linear predictors are

$$\beta^* = \mathrm{VAR}(X)^{-1}\mathrm{COV}(X,Y), \qquad \beta_0^* = \mathbb{E}Y - \mathbb{E}X^\top \beta^*.$$

We have to assume that $\mathbb{E}X_i^2 < \infty$ for $i = 1, \dots, d$, $\mathbb{E}Y^2 < \infty$ and the variance-covariance matrix $\mathrm{VAR}(X) = (\mathrm{Cov}(X_i, X_j), i, j = 1, \dots, d)$ is positive definite (nonsingular). △

1.1.7 EXAMPLE (Quantile Regression). Instead of the quadratic loss funcion, in the model of regression we can use the loss defined by (1.1.4). Let $Z = (X, Y)$, where X is a d-dimensional random vector and Y is one-domensional random variable. The *quantile regression* obtains if we put

$$\rho(\beta_0, \beta, x, y) = \begin{cases} (\beta_0 + x^\top \beta - y)(1-p) & \text{for } y \le \beta_0 + x^\top \beta; \\ (y - \beta_0 - x^\top \beta)p & \text{for } y > \beta_0 + x^\top \beta. \end{cases}$$

The name 'quantile regression' is explained as follows. If we assume that $Y = \beta_0^* + X^\top \beta^* + W$, with W independent of X, then $\beta_0^* + x^\top \beta^*$ is the pth quantile of the conditional distribution of Y given $X = x$. In general, there is no closed explicit expression for (β_0^*, β^*). In particular, for $p = 1/2$ we obtain *Least Absolute Deviations* regression (LAD, known also as MAD – '*Minimum* Absolute Deviations'). Indeed, the quantile loss can be the equivalently defined as

$$\rho(\beta_0, \beta, x, y) = |y - \beta_0 - x^\top \beta|.$$

△

1.1.8 EXAMPLE (Linear Classification). Let $Z = (X, K)$, where $X = (X_1, \dots X_d)^\top$ is a random vector and K is a random variable with two possible values. Variable K is interpreted as a label of class to which an "object" belongs and $X = (X_1, \dots X_d)^\top$ is a vector of observed "features" of this object. The problem is analogous as in the regression models: we are to predict K, given X – that is to "classify" the object. It is convenient to choose $\{-1, 1\}$ as the set of values of K. As in previous examples related to regression, we restrict attention to classification based on linear functions $\beta_0 + x^\top \beta$. Suppose that the predicted label is $\hat{K} = \operatorname{sign}(\beta_0 + X^\top \beta)$. The most natural loss function is the indicator of the incorrect guess:

$$\rho(\beta_0, \beta, x, k) = \mathbb{1}(\operatorname{sign}(\beta_0 + x^\top \beta \neq k).$$

For this $0-1$ loss, the risk is equal to $\mathbb{P}(\hat{K} \neq K)$, the probability of misclassification. However, the $0-1$ loss has its disadvantages. It leads to nonconvex minimization problems, multimodality of the risk function and high computational complexity. There is a loss function which is to some extent 'similar', but much more concenient, named hinge loss. It is defined by

(1.1.9)

$$\rho(\beta_0, \beta, x, k) = \big(1 - k(\beta_0 + x^\top \beta)\big)_+ = \begin{cases} 1 - \beta_0 - x^\top \beta & \text{if } k = 1 \text{ and } \beta_0 + x^\top \beta < 1; \\ 1 + \beta_0 + x^\top \beta & \text{if } k = -1 \text{ and } \beta_0 + x^\top \beta > -1; \\ 0 & \text{otherwise.} \end{cases}$$

The hinge loss is convex in (β_0, β) and computationally friendly. △

1.1.10 EXAMPLE (Cumulant Generating Function). Let Z be one-dimensional random variable and consider the following 'LINEX' loss function:

$$\rho(a,z) = \exp\{\kappa(z-a)\} - \kappa(z-a) - 1,$$

where $\kappa > 0$. The corresponding M-functional is $a_* = \frac{1}{\kappa}\log \mathbb{E}\exp\{\kappa Z\}$.

$\triangle$

1.2 M-estimators and asymptotic results

We consider function $R : \mathbb{R}^d \to \mathbb{R}$ given by

$$R(a) = \mathbb{E}\rho(a, Z).$$

The sample analogue of R is

$$R_n(a) = \frac{1}{n}\sum_{i=1}^{n} \rho(a, Z_i),$$

where random vectors $Z_1, \ldots, Z_n$ are independent and identically distributed. Let a_* and a_n denote minimizers of $R(a)$ and $R_n(a)$, respectively. We say a_n is an M-estimator.

The main issues is to prove *consistency* and *asymptotic normality*. There are many results which establish these properties of M-estimators, under different sets of assumptions. I have chosen to focus on *convex* loss functions. This is because M-estimators defined by convex minimization are computationally friendly, and there are no problems with multimodality. And, this class of M-estimators was the topic of my past work.

Our standing assumption will be the following.

1.2.1 Assumption. *1) $\rho(a,z)$ is convex as a function of $a \in \mathbb{R}^d$, for every fixed $z \in \mathcal{Z}$.*

2) $R(a)$ is well defined (the expectation exists) for every $a \in \mathbb{R}^d$.

Convexity plays a crucial role in our proofs. In the proofs we need the following lemma, taken from Niemiro (1992), which is an easy consequence of standard results on convex functions (Rockafellar 1970).

1.2.2 Lemma. *Let $R_n(a)$, $n = 1, \ldots$ be convex random functions on $\mathbb{R}^d$. Assume that $R(a)$ is a random function such that $R_n(a) \to R(a)$ (a) in probability (b) almost surely, for each fixed a. Then on each compact $K \subset \mathbb{R}^d$ we have uniform convergence:*

$$\sup_{a\in K} |R_n(a) - R(a)| \longrightarrow 0.$$

(a) in probability (b) almost surely, respectively.

Consistency of M-estimators

1.2.3 Theorem. *If a_* is the unique minimizer of $R(a)$ then $a_n \to a_*$ with probability 1.*

Note that under the assumptions of Theorem 1.2.3, a point a_n which minimizes $R_n(a)$ exists almost surely, at least for sufficiently large n. It may be not unique, but then we can choose a_n arbitrarily, subject to the condition that selection is measurable. These facts will be easily seen in the proof. The existence of a measurable selector is shown in Niemiro (1992, Appendix).

Proof of Theorem 1.2.3. Continuity of R and the fact that a_* is its unique minimizer imply that for arbitrary $\varepsilon > 0$ there exists $\tau > 0$ such that $R(a) > R(a_*) + 2\tau$ for $|\,a - a_*| = \varepsilon$. By SLLN we have $|R_n(a) - R(a)| \longrightarrow_{\text{a.s.}} 0$ for every fixed a. Lemma 1.2.2 implies that $\sup_{|a-a_*|\le\varepsilon} |R_n(a) - R(a)| \longrightarrow_{\text{a.s.}} 0$. Then the following inequalities hold with probability one for sufficiently large n:

$$R_n(a) > R(a) - \tau > R(a_*) + \tau \qquad \text{for } |\,a - a_*| = \varepsilon,$$

$$R_n(a_*) < R(a_*) + \tau.$$

Summarizing, $R_n(a) > R_n(a_*)$ for every a such that $|a - a_*| = \varepsilon$. By convexity of R_n we get $|\,a_n - a_*| < \varepsilon$. □

Asymptotic normality of M-estimators

To study asymptotic normality of M-estimators defined by a convex loss, we need the notion of *subgradient*. Let $\psi(a, z) \in \mathbb{R}^d$ be a subgradient of convex function $\rho(a, z)$. By definition, this means that for all $a_1, a_2 \in \mathbb{R}^d$ we have

$$\rho(a_2, z) - \rho(a_1, z) \ge \psi(a_1, z)^\top (a_2 - a_1). \tag{1.2.4}$$

A subgradient ψ of a convex function ρ always exists, but it may be not unique. Actually it is not unique in the case when ρ is defined as (1.1.9) or (1.1.4). Nevertheless we will write $\psi(a, z) = \partial\rho(a, z)$, understanding that ψ is some (measurable selection) of subgradient. We refer again to Rockafellar (1970) and Niemiro (1992, Appendix) for details.

Let $\nabla R(a)$ and $D = \nabla\nabla^\top R(a)$ denote gradient and matrix of second partial derivatives of $R(a)$, respectively.

1.2.5 Theorem. *Assume that*

1) $R(a)$ is twice differentiable at a_* and matrix $D = \nabla\nabla^\top R(a_*)$ is (strictly) positive definite,

2) there exists $\eta > 0$ such that $\mathbb{E}\,|\partial\rho(a, Z)|^2 < \infty$ for all a satisfying $|a - a_*| < \eta$.

Then

$$\sqrt{n}(a_n - a_*) \longrightarrow_d \mathrm{N}(0, D^{-1}VD^{-1}),$$

where

$$V = \mathrm{VAR}(\partial\rho(a_*, Z)).$$

To underline the main ideas behind asymptotic normality of M-estimators, we first give a simple heuristic argument which explains the conclusion of our Theorem 1.2.5. A rigorous proof in deferred to the next subsection (*optional), because it is rather technical and makes essential use of special properties of convex functions.

Sketch of the proof of Theorem 1.2.5. The following approximate equation is a (very short) Taylor expansion of $R_n(a) - R(a)$ up to linear term:

$$R_n(a) - R(a) \simeq R_n(a_*) - R(a_*) + (\partial R_n(a_*) - \nabla R(a_*))^\top (a - a_*)$$

Notice that $\nabla R(a_*) = 0$, so this term can be dropped. Now let us write $\alpha = \sqrt{n}(a - a_*)$ so that $a = a_* + \alpha/\sqrt{n}$. Reparametrize the last displayed equation in terms of α, multiply by n and rearrange. Then use a (genuine) Taylor expansion $nR(a_* + \alpha/\sqrt{n}) - nR(a_*) \simeq \alpha^\top D\alpha/2$ and write $G_i = \partial\rho(a_*, Z_i)$ to obtain

$$\begin{aligned} nR_n\left(a_* + \frac{\alpha}{\sqrt{n}}\right) - nR_n(a_*) &\simeq nR\left(a_* + \frac{\alpha}{\sqrt{n}}\right) - nR(a_*) + \sqrt{n}\partial R_n(a_*)^\top\alpha \\ &\simeq \frac{1}{2}\alpha^\top D\alpha + \frac{1}{\sqrt{n}}\sum_{i=1}^n G_i^\top\alpha. \end{aligned}$$

The RHS of this approximate equality is a quadratic function with a deterministic quadratic term. The linear term on the RHS is a sum of i.i.d. random vectors G_i, properly standardized, and consequently going to $\mathrm{N}(0, V)$. The quadratic function on the RHS has the minimum at $\tilde{\alpha}_n = -D^{-1}\sum_{i=1}^n G_i/\sqrt{n}$. It is therefore plausible that the minimum of the LHS denoted by α_n is close to $\tilde{\alpha}_n$. We know that $\tilde{\alpha}_n \to_\mathrm{d} \mathrm{N}(0, D^{-1}VD^{-1})$ and we expect that α_n has the same limiting distribution. If we return to the initial parameter a, we see that $\alpha_n = \sqrt{n}(a_n - a_*)$. □

1.2.6 REMARK. There are many theorems with conclusions similar to our Theorems 1.2.3 i 1.2.5. In general it is not necessary to assume convexity of ρ, but then other and rather strong assumptions have to be made. In some versions of asymptotic normality, D is defined as

$\mathbb{E}\nabla\nabla^\top\varrho(a_*, Z)$, instead of $\nabla\nabla^\top\mathbb{E}\varrho(a_*, Z)$ as in our Theorem 1.2.5. We have chosen convexity as our standing assumption because it simplifies the proofs and is fulfilled in many interesting xmplles.

1.2.7 EXAMPLE (Sample Quantiles). Consider the loss function which defines a p-quantile as M-functional, Example 1.1.2 or (1.1.4). The corresponding M-estimator is, of course, the pth sample quantile. If we assume that random variable Z has the probability density f which is continuous at a_*, the unique p-quantile, and $f(a_*) > 0$ then the assumptions of Theorem 1.2.5 are fulfilled. The sample quantile a_n is asymptotically normal,

$$\sqrt{n}(a_n - a_*) \longrightarrow_d \mathrm{N}\left(0, \frac{p(1-p)}{f(a_*)^2}\right).$$

This example shows that the assumption 1) of Theorem 1.2.5 can well be satisfied even if the loss ρ is not differentiable. Although $\mathbb{E}(\mathrm{d}^2/\mathrm{d}a^2)\rho(a, z)$ does not exist for ρ defined by (1.1.4), $(\mathrm{d}^2/\mathrm{d}a^2)\mathbb{E}\rho(a, z)$ is well defined and positive. △

* A rigorous proof of asymptotic normality

In the proof of asymptotic normality we will need some useful properties of subgradient, listed below. If $\psi(a, z) = \partial\rho(a, z)$ then

$$\begin{aligned} 0 &\le \rho(a_2, z) - \rho(a_1, z) - (a_2 - a_1)^\top\psi(a_1, z) \\ &\le [\psi(a_2, z) - \psi(a_1, z)]^\top(a_2 - a_1). \end{aligned} \tag{1.2.8}$$

For every fixed $e \in \mathbb{R}^d$,

$$[\psi(a + te, z) - \psi(a, z)]^\top e \quad \text{is increasing in } t \in \mathbb{R}. \tag{1.2.9}$$

Proof of Theorem 1.2.5. We will just translate the approximate equations to precise statements. Write $\psi(a, Z) = \partial\rho(a, Z)$. To begin with, note that

$$\mathbb{E}\psi(a, Z) = \nabla R(a)$$

at each point a of differentiability of R. Indeed, the definition of subgradient implies that for every $e \in \mathbb{R}^d$ and $t > 0$,

$$-\frac{\rho(a - te, Z) - \rho(a, Z)}{t} \le \psi(a, Z)^\top e \le \frac{\rho(a + te, Z) - \rho(a, Z)}{t}.$$

It is now enough to take expectations and pass to the limit with $t \to 0$. We shall prove that for every fixed α,

$$\begin{aligned} &nR_n\left(a_* + \frac{\alpha}{\sqrt{n}}\right) - nR_n(a_*) \\ &- nR\left(a_* + \frac{\alpha}{\sqrt{n}}\right) + nR(a_*) - \sqrt{n}\partial R_n(a_*)^\top\alpha \to_{\mathrm{p}} 0 \end{aligned} \tag{1.2.10}$$

Consider random variables

$$H_{ni} = \rho\left(a_* + \frac{\alpha}{\sqrt{n}}, Z_i\right) - \rho(a_*, Z_i) - \psi(a_*, Z_i)^\top \frac{\alpha}{\sqrt{n}}.$$

We have

$$\sum_{i=1}^n H_{ni} = nR_n\left(a_* + \frac{\alpha}{\sqrt{n}}\right) - nR_n(a_*) - \sqrt{n}\partial R_n(a_*)^\top \alpha$$

and

$$n\mathbb{E}H_{ni} = nR\left(a_* + \frac{\alpha}{\sqrt{n}}\right) - nR(a_*).$$

Since $H_{n1}, \ldots, H_{ni}$ are i.i.d., we have $\operatorname{Var}\sum_{i=1}^n H_{ni} = \sum_{i=1}^n \operatorname{Var} H_{ni} \le n\mathbb{E}H_n^2$, where H_n is defined exactly as H_{ni} but with Z_i replaced by $Z =_{\mathrm{d}} Z_i$. Now use (1.2.8) to obtain $nH_n^2 \le T_n^2$, where

$$T_n = \left(\psi\left(a_* + \frac{\alpha}{\sqrt{n}}, Z\right) - \psi(a_*, Z)\right)^\top \alpha.$$

Random variables are nonnegative and monotonically decrease, by (1.2.8). Thus $T_n \searrow T \ge 0$. We have $\mathbb{E}T_n = \nabla R\left(a_* + \alpha/\sqrt{n}\right) - \nabla R(a_*) \to 0$, because ∇R is continuous at a_*. Moreover, T_n have finite second moments, so by Lebesgue dominated convergence we infer that $\mathbb{E}T = 0$, so $\mathbb{P}(T = 0) = 1$, and again by dominated convergence $\mathbb{E}T_n^2 \to T^2 = 0$. Therefore $\sum_{i=1}^n \mathbb{E}H_{ni}^2 \to 0$ and by Chebyshev inequality we get $\sum_{i=1}^n (H_{ni} - \mathbb{E}H_{ni}) \to_{\mathrm{p}} 0$, which is just (1.2.10).

From (1.2.10) and the Taylor expansion $R(a) - R(a_*) = (a - a_*)^\top D(a - a_*)/2 + o(|a - a_*|^2)$ it follows that

$$\begin{aligned} &nR_n\left(a_* + \frac{\alpha}{\sqrt{n}}\right) - nR_n(a_*) \\ &\quad - \frac{1}{2}\alpha^\top D\alpha - \sqrt{n}\partial R_n(a_*)^\top \alpha \to_{\mathrm{p}} 0 \end{aligned} \tag{1.2.11}$$

The next step is to strengthen pointwise convergence in (1.2.11) to uniform convergence on compacts. It is enough to invoke Lemma 1.2.2 to obtain

$$\begin{aligned} \sup_{|\alpha| \le M} \Bigg| &nR_n\left(a_* + \frac{\alpha}{\sqrt{n}}\right) - nR_n(a_*) \\ &- \frac{1}{2}\alpha^\top D\alpha - \sqrt{n}\partial R_n(a_*)^\top \alpha \Bigg| \to_{\mathrm{p}} 0. \end{aligned} \tag{1.2.12}$$

Finally, rewrite (1.2.12) as $\sup_{|\alpha| \le M} |q_n(\alpha) - \tilde{q}_n(\alpha)| \to_{\mathrm{p}} 0$, where $q_n(\alpha) = nR_n\left(a_* + \alpha/\sqrt{n}\right) - nR_n(a_*)$ has a minimum at α_n and the quadratic function $\tilde{q}_n(\alpha) = (\alpha - \tilde{\alpha}_n)^\top D(\alpha - \tilde{\alpha}_n)/2 - \tilde{\alpha}_n^\top \tilde{\alpha}_n/2$ has the minimum at $\tilde{\alpha}_n = -D^{-1}\sum_{i=1}^n G_i/\sqrt{n}$, with $G_i = \psi(a_*, Z_i)$. Let $2\tau = \inf_{|e|=\varepsilon} e^\top De$. If $q_n(\tilde{\alpha}_n) < \tilde{q}_n(\tilde{\alpha}_n) + \tau$ and $q_n(\alpha) > \tilde{q}_n(\alpha) - \tau$ for every α such that $|\alpha - \tilde{\alpha}_n| = \varepsilon$ then the convex function q_n has the value at $\tilde{\alpha}_n$ less than its values on the sphere with radius ε and centre $\tilde{\alpha}_n$. Consequently, its minimum α_n must lie inside the ball around $\tilde{\alpha}_n$. This happens with probability approaching one. Thus $|\tilde{\alpha} - \alpha_n| \to_{\mathrm{p}} 0$ and α_n must have the same limit in distribution as $\tilde{\alpha}_n$, namely $\mathrm{N}(0, D^{-1}VD^{-1})$. □

1.3 Maximum likelihood

Cosider another example of a loss function, which is in fact much more than just an example.

1.3.1 EXAMPLE (Log-likelihood). Consider a parametric family of probability densities $\{p_\theta(\cdot) : \theta \in \Theta\}$ on a sample space $\mathcal{X}$. Let

$$\rho(\theta, x) = -\log p_\theta(x).$$

Minimization of the empirical loss $R_n(\theta) = -\sum_{i=1}^n \log p_\theta(X_i)/n$ is equivalent to computing the Maximum Likelihood Estimate (MLE). If random variable X has probability distribution $p(\cdot)$ (for simplicity it will be identified with a density) then a minimum of $R(\theta) = -\mathbb{E}\log p_\theta(X)$ is a point θ_* which minimizes the Kullback-Leibler divergence $D(p\|p_\theta)$. We say p_{θ_*} is the information projection of p onto the parametric family $\{p_\theta\}$. △

Recall the definition of the Kullback-Leibler divergence. If P and Q are probability distributions on $\mathcal{X}$ which are mutually absolutely continuous ($Q \ll P$ and $Q \ll P$) then

$$\mathcal{D}(P\|Q) = \mathbb{E}_{X\sim P} \log \frac{\mathrm{d}P}{\mathrm{d}Q}(X) = \int_{\mathcal{X}} \log \frac{\mathrm{d}P}{\mathrm{d}Q}(x) P(\mathrm{d}x),$$

where $\dfrac{\mathrm{d}P}{\mathrm{d}Q}$ is the Radon-Nikodym derivative. By Jensen inequality,

$$\mathcal{D}(P\|Q) = \int_{\mathcal{X}} -\log \frac{\mathrm{d}Q}{\mathrm{d}P}(x) P(\mathrm{d}x) \geq -\log \int_{\mathcal{X}} \frac{\mathrm{d}Q}{\mathrm{d}P}(x) P(\mathrm{d}x) - \log \int_{\mathcal{X}} Q(\mathrm{d}x) = -\log 1 = 0.$$

We obtain the following basic fact:

$$\mathcal{D}(P\|Q) \geq 0.$$

Moreover, since $-\log$ is strictly convex, the equality in Jensen inequality obtains only if $\dfrac{\mathrm{d}Q}{\mathrm{d}P}(x) \equiv 1$ almost everywhere. Thus

$$\mathcal{D}(P\|Q) = 0 \quad \text{if and only if} \quad P = Q.$$

If P and Q have densities p and q with respect to some σ-finite measure, the definition ofthe Kullback-Leibler divergence can be rewritten as follows. To simplify notation, let us introduce some conventions. The reference measure will be denoted by $\mathrm{d}x$ (although it is not necessarily the Lebesgue measure), so we write $P(\mathrm{d}x) = p(x)\mathrm{d}x$ and $Q(\mathrm{d}x) = q(x)\mathrm{d}x$. Moreover, we will identify probability measures with their densities. If $\{x : p(x) > 0\} = \{x : q(x) > 0\}$ almost everywhere $[\mathrm{d}x]$ then

$$\mathcal{D}(p\|q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x.$$

Continuing Example 1.3.1, asymptotic properties of MLE are obtained as special case of those for M-estimators.

1.3.2 Corollary. *Consider a family of densities* $\{p_\theta : \theta \in \Theta\}$, *where* $\Theta \subseteq \mathbb{R}^d$. *Let* $X_1, \dots, X_n, \dots$ *be a sequence of i.i.d. random variables with density* p, *not necessarily belonging to the parametric family. Assume that the information projection* p_{θ_*} *exists and is unique. If* θ_n *is the MLE based on* $X_1, \dots, X_n$ *(sample of size* n*). If the conclusion of Theorem 1.2.3 holds then*

$$\theta_n \to \theta_* \quad \textit{almost surely.}$$

If p *belongs to the parametric family, that is* $p = p_{\theta_0}$ *for some* $\theta_0 \in \Theta$ *(and* $p \neq p_\theta$ *for* $\theta \neq \theta_0$*), then we obtain strong consistency of MLE:*

$$\theta_n \to \theta_0 \quad \textit{almost surely.}$$

There are two points that should be emphasized about Corollary 1.3.2. First, this result explains what happens if the parametric model is *misspecified*. It is arguably not realistic to assume that a model *exactly* describes the random phenomenon at hand. The second point is that the main assumption of Theorem 1.2.3 is convexity of loss function. The log-likelihood $-\log p_\theta(x)$ in general need not be convex in θ (although it is for *some* interesting paramertric families). However, as we noted earlier, the conclusion of Theorem 1.2.3 remains to be true without convexity, if some other "regularity assumptions" are fulfilled. The convergence of MLE then holds. An example is given in Problem 6.

Now we proceed to the asymptotic normality of MLE. In the following Theorem we skip technical assumptions (which are neither simple nor easy) and focus on the main ideas.

1.3.3 Corollary. *Consider a situation described in Theorem 1.3.2. Additionally assume that the parametric family satisfies regularity assumptions which permit differentiating* $-\log p_\theta(X)$ *twice and interchanging the order of operators* ∇_θ *and* $\int \cdots \mathrm{d}x$. *If the conclusion of Theorem 1.2.5 holds then*

$$\sqrt{n}(\theta_n - \theta_*) \to \mathrm{N}(0, D^{-1} V D^{-1}) \quad \textit{in distribution,}$$

where

$$V = \mathrm{VAR}(\nabla \log p_{\theta_*}(X)) = \mathbb{E} \nabla \log p_{\theta_*}(X) \nabla^\top \log p_{\theta_*}(X),$$
$$D = -\nabla \nabla^\top \mathbb{E} \log p_{\theta_*}(X) = -\mathbb{E}(\nabla \nabla^\top \log p_{\theta_*}(X)).$$

If p *belongs to the parametric family, that is* $p = p_{\theta_0}$ *for some* $\theta_0 \in \Theta$, *then*

$$\sqrt{n}(\theta_n - \theta_0) \to \mathrm{N}(0, I^{-1}(\theta_0)) \quad \textit{in distribution,}$$

where $I(\theta)$ *is the Fisher information matrix.*

Proof. The first conclusion immediate follows from Theorem 1.2.5. To obtain the second

conclusion, we assume that $p = p_{\theta_0}$ and show that then $V = D = I(\theta_0)$. Indeed,

$$\begin{aligned}
-D = \mathbb{E}(\nabla\nabla^\top \log p_{\theta_0}(X)) &= \int p(x)\nabla\nabla^\top \log p_{\theta_0}(x)\mathrm{d}x \\
&= \int p(x)\frac{p_{\theta_0}(x)\nabla\nabla^\top p_{\theta_0}(x) - \nabla p_{\theta_0}(x)\nabla^\top p_{\theta_0}(x)}{p_{\theta_0}(x)^2}\mathrm{d}x \\
\text{(Assumtion } p = p_{\theta_0} \text{ is used here)} \quad &= \int \frac{p_{\theta_0}(x)\nabla\nabla^\top p_{\theta_0}(x) - \nabla p_{\theta_0}(x)\nabla^\top p_{\theta_0}(x)}{p_{\theta_0}(x)}\mathrm{d}x \\
&= \int \nabla\nabla^\top p_{\theta_0}(x)\mathrm{d}x - \int \frac{\nabla p_{\theta_0}(x)}{p_{\theta_0}(x)}\frac{\nabla^\top p_{\theta_0}(x)}{p_{\theta_0}(x)}\mathrm{d}x \\
&= \nabla\nabla^\top \int p_{\theta_0}(x)\mathrm{d}x - \int \nabla \log p_{\theta_0}(x)\nabla^\top \log p_{\theta_0}(x)\mathrm{d}x \\
&= 0 - V.
\end{aligned}$$

The above calculation is standard. Either side of the equation (V or D) can serve as the definition of $I(\theta_0)$. □

If the model is misspecified then in general $V \neq D$ and neither is equal to $I(\theta_*)$. An example is given in Problem 6. Note in passing that $I(\theta_*)$ can be consistently estimated by $I(\theta_n)$ but estimation of D and V is more difficult.

Problems

1. Give a full and rigorous derivation of all the results stated in Examples 1.1.2, 1.1.5, 1.1.6, 1.1.7 and 1.1.10. The following auxillary facts may be useful in Example 1.1.2.

Lemma. Let $F : \mathbb{R} \to \mathbb{R}$ be a nondecreasing and right-continuous function. For a fixed, define function I by $I(b) = \int_a^b F(x)\mathrm{d}x$. Show that I is convex and for every $b > a$ there exist one-sided derivatives

$$I'_-(b) = F(b-) \leq I'_+(b) = F(b) = F(b+).$$

We use the notation $F(b-) = \lim_{x\nearrow} F(x)$ and the one-sided derivatives are $I'_-(b) = \lim_{x\nearrow} \frac{F(x) - F(b)}{x - b}$ and $I'_+(b) = \lim_{x\searrow} \frac{F(x) - F(b)}{x - b}$.

Lemma. Let Z be a one-dimensional random variable. Show that

$$\mathbb{E}(Z - a)\mathbb{1}(Z > a) = \int_a^\infty \mathbb{P}(Z > z)\mathrm{d}z.$$

2. Prove the following property of the M-functional defined via the *hinge* loss for linear classification, Example 1.1.8. If we consider an affine nonsingular transformation $\mathbb{R}^d \to \mathbb{R}^d$ of the feature vector: $\tilde{X} = T \cdot X + \alpha$, where T is a $d \times d$ nonsingular matrix and $\alpha \in \mathbb{R}^d$. Show that a minimizer of $Q(\beta_0, \beta)$ is equivariant: $\tilde{\beta}_0^* + \tilde{X}^\top \tilde{\beta}^* = \beta_0^* + X^\top \beta^*$.

   Show that choosing the "margin width" to be 1 in (1.1.9) is unimportant. If we alter the definition of the hinge loss to $\rho(\beta_0, \beta, x, k) = \big(h - k(\beta_0 + x^\top \beta)\big)_+$ with arbitrarily chosen $h > 0$, then the results of classification will be the same.

3. Prove (1.2.8) and (1.2.9).

4. Compute the Kullback-Leibler divergence between two binomial distributions: for $p_0, p_1 \in]0,1[$, compute $\mathcal{D}(\mathrm{Bin}(n, p_0) \| \mathrm{Bin}(n, p_1))$.

5. Compute the Kullback-Leibler divergence between two Poisson distributions: for $\lambda_0, \lambda_1 \in]0, \infty[$, compute $\mathcal{D}(\mathrm{Poiss}(\lambda_0) \| \mathrm{Poiss}(\lambda_1))$.

6. Compute the Kullback-Leibler divergence between two normal distributions: compute $\mathcal{D}(\mathrm{N}(\mu_0, \sigma_0^2) \| \mathrm{N}(\mu_1, \sigma_1^2))$.

   Find the information projection of $\mathrm{N}(\mu_0, \sigma_0^2)$ on the one-parameter family of distributions $\{\mathrm{N}(\mu, \sigma_1^2) : \mu \in \mathbb{R}\}$, where $\sigma_1 > 0$ is fixed.

   Find the information projection of $\mathrm{N}(\mu_0, \sigma_0^2)$ on the one-parameter family of distributions $\{\mathrm{N}(\mu_1, \sigma^2) : \sigma^2 \in]0, \infty[\}$, where μ_1 is fixed.

   Compute the asymptotic variance of the MLE of σ in the following situation (model misspecification). Assume that the sample $X_1, \ldots, X_n$ is drawn from $\mathrm{N}(\mu_0, \sigma_0^2)$, but we wrongly assume that the probability distribution belongs to the family $\{\mathrm{N}(\mu_1, \sigma^2) : \sigma^2 \in]0, \infty[\}$ and we compute the MLE $\hat{\sigma}_n^2$ under this assumption. Show that the MLE is asymptotially normal: $\sqrt{n}(\hat{\sigma}_n^2 - \sigma_*^2) \to \mathrm{N}(0, s_{\mathrm{as}}^2)$. Compute σ_*^2 and s_{as}^2. Discuss the special case $\mu_1 = \mu_0$.

   Hint: It is more convenient to use σ^2 and not σ as a parameter. The answer is the following: $\sqrt{n}(\hat{\sigma}_n^2 - (\sigma_0^2 + (\mu_1 - \mu_0)^2)) \to \mathrm{N}(0, 4(\mu_1 - \mu_0)^2\sigma_0^2 + 2\sigma_0^4)$. This result can be derived directly, using only the CLT and formulas for the moments of normal ditribution. The result is consistent with the conclusion of Theorem 1.2.5, despite the fact that the log-likelihood is *not* concave in σ^2. In this example we have $V = (4(\mu_1 - \mu_0)^2\sigma_0^2 + 2\sigma_0^4)/(\sigma_0^2 + (\mu_1 - \mu_0)^2)^4$ and $D = -1/((\sigma_0^2 + (\mu_1 - \mu_0)^2)^2$.

7. Consider a one-parameter family of probability distributions (densities) $\{p_\theta : \theta \in \mathbb{R}\}$. Show that for every θ_0 we have $\dfrac{\mathrm{d}^2}{\mathrm{d}\theta^2}\mathcal{D}(p_{\theta_0} \| p_\theta)_{|\theta=\theta_0} = I(\theta_0)$, where $I(\theta_0)$ is the *Fisher information*. Show also that $\dfrac{\mathrm{d}^2}{\mathrm{d}\theta^2}\mathcal{D}(p_\theta \| p_{\theta_0})_{|\theta=\theta_0} = I(\theta_0)$.

Chapter 2

Bayesian Model

From the methodological perspective, the Bayesian approach boils down to modelling uncertainty consistently in terms of probability. From the mathematical perspective, Bayesian statistics is a branch of probability theory focussed on the concept of conditioning.

2.1 Fubini and Bayes

A Bayesian model is constructed from a family of conditional distributions $\{P_\theta(\mathrm{d}x) : \theta \in \Theta\}$ on a sample space $\mathcal{X}$ and a probability distribution (a "prior') $\Pi(\mathrm{d}\theta)$ on Θ. This is done via a (slightly generalized) *Fubini* theorem. *Bayes* theorem can be regarded as the inverse: we decompose the joint distribution into a marginal $P(\mathrm{d}x)$ and the family of conditional distributions $\Pi_x(\mathrm{d}\theta)$ ("posteriors"). In a nutshell,

$$P_\theta(\mathrm{d}x)\Pi(\mathrm{d}\theta) = \Pi_x(\mathrm{d}\theta)P(\mathrm{d}x).$$

Below we present details. In these notes, the conditional expectation and conditional probability is treated at an intermediate level of abstraction (and with moderate rigor). To avoid unpleasant pathologies we will restrict considerations to Polish spaces. As customary in the statistical literature, we use the canonical probability spaces, so that there is no unnecessary distinction between 'probability' and 'probability distribution'.

Let $\mathcal{X}$ and Θ be Polish spaces equipped with their Borel σ-fields $\mathfrak{X}$ and $\mathfrak{Q}$. Let us begin with the definition of a transition kernel. A function $P : \mathcal{X} \times \mathfrak{Q} \to [0, 1]$ is called a *transition kernel*, if

- For every $\theta \in \Theta$, $P_\theta(\cdot)$ is a probability distribution on $(\mathcal{X}, \mathfrak{X})$.
- For every $B \in \mathfrak{X}$, function $P_\cdot(B)$ is $\mathfrak{Q} \to \mathfrak{B}(\mathbb{R})$-measurable.

Note that we use the notation $P_\theta(B)$, where $\theta \in \Theta$ and $B \subseteq \mathcal{X}$, $B \in \mathfrak{X}$. We will say that P is a transition probability '$\Theta \to \mathcal{X}$' (the respective σ-fields will not be explicitly mentioned, because we always assume they are Borel σ-fields in the corresponding spaces.

2.1.1 Theorem (Fubini). *If $\{P_\theta(\mathrm{d}x) : \theta \in \Theta\}$ is a transition kernel $\Theta \to \mathcal{X}$ and $\Pi(\mathrm{d}\theta)$ is a probability distribution on Θ then there exists a probability measure $\mathbb{P}$ on $\Omega = \Theta \times \mathcal{X}$ such that*

$$\mathbb{P}(C \times B) = \int_C P_\theta(B)\Pi(\mathrm{d}\theta)$$

for every (Borel) $B \subseteq \mathcal{X}$ and every (Borel) $C \subseteq \Theta$. More generally,

$$\iint_\Omega h(\theta, x)\mathbb{P}(\mathrm{d}\theta, \mathrm{d}x) = \int_\Theta \int_\mathcal{X} h(\theta, x)P_\theta(\mathrm{d}x)\Pi(\mathrm{d}\theta)$$

for every nonnegative or bounded (and measurable) function $h : \Omega \to \mathbb{R}$.

Having constructed the probability measure on $\Omega = \Theta \times \mathcal{X}$, we can treat the coordinates θ and x as *random variables*. Formally, we define $\vartheta : \Omega \to \Theta$ and $X : \Omega \to \mathcal{X}$ by

$$\vartheta(\theta, x) = \theta, \qquad X(\theta, x) = x.$$

Thus we obtain random variables ϑ and X on the *canonical* probability space $(\Omega, \mathbb{P}, \mathfrak{Q} \otimes \mathfrak{X})$. 'Canonical' space Ω is the set of possible values of all (two in this example) variables of interest. Of course, Π becomes the probability distribution of ϑ and P_θ is a regular version of the probability distribution of X given $\vartheta = \theta$:

$$\mathbb{P}(\vartheta \in \mathrm{d}\theta) = \Pi(\mathrm{d}\theta), \quad P_\theta(\mathrm{d}x) = \mathbb{P}(X \in \mathrm{d}x | \vartheta = \theta).$$

The marginal probability distribution of X is the measure on $\mathcal{X}$ given by

$$P(\mathrm{d}x) = \int P_\theta(\mathrm{d}x)\Pi(\mathrm{d}\theta).$$

The next step is to decompose the joint probability distribution "in the opposite direction". This is an abstract version of the Bayes theorem.

2.1.2 Theorem (Bayes). *If $\mathbb{P}$ is a probability measure on $\Omega = \Theta \times \mathcal{X}$ then there exists a transition kernel $\mathcal{X} \to \Theta$ denoted $\Pi_x(\mathrm{d}\theta)$ such that*

$$\mathbb{P}(C \times B) = \int_B \Pi_x(C)P(\mathrm{d}x)$$

for every (Borel) $B \subseteq \mathcal{X}$ and every (Borel) $C \subseteq \Theta$. Measure P is the marginal distribution on $\mathcal{X}$, that is $P(B) = \mathbb{P}(\Theta \times B)$. More generally,

$$\iint_\Omega h(\theta, x)\mathbb{P}(\mathrm{d}\theta, \mathrm{d}x) = \int_\mathcal{X} \int_\Theta h(\theta, x)\Pi_x(\mathrm{d}\theta)P(\mathrm{d}x)$$

for every nonnegative or bounded (and measurable) function $h : \Omega \to \mathbb{R}$.

Of course, $\Pi_x(\mathrm{d}\theta)$ is a regular version of the conditional distribution:

$$\Pi_x(\mathrm{d}\theta) = \mathbb{P}(\vartheta \in \mathrm{d}\theta | X = x).$$

We say $\Pi(\mathrm{d}\theta)$ is the *prior* distribution and $\Pi_x(\mathrm{d}\theta)$ is the *posterior* distribution. In principle, the prior should describe our uncertainty about the value of θ before observing $X = x$. Upon observing $X = x$, we update Π to Π_x. The information contained in data x is used to replace unconditional distribution (prior) by the conditional one (posterior). This is the essence of the Bayesian statistical inference.

Most often, probability distributions in statistical models are defined via densities with respect to some "natural" measures. A standard convention in Bayesian statistics is to identify probability distributions with their densities. Let us now rewrite Theorems 2.1.1 and 2.1.2 in a more familiar and elementary way, in terms of densities. We begin with a family of densities $\{p_\theta\}$ on $\mathcal{X}$ parametrized by $\theta \in \Theta$ (all these densities are with respect to a measure denoted by $\mathrm{d}x$). On the space Θ we have density π with respect to a measure denoted by $\mathrm{d}\theta$. A joint density on $\Theta \times \mathcal{X}$ is given by $p(\theta, x) = p_\theta(x)\pi(\theta)$. According to the definition of conditional density we can write $p_\theta(x) = p(x|\theta)$ and obtain immediately the famous *Bayes formula*:

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)}, \quad \text{where} \quad p(x) = \int p(x|\theta)\pi(\theta)\mathrm{d}\theta. \tag{2.1.3}$$

Let us repeat the basic "vocabulary" of Bayesian statistics.

- $p(x|\theta)$ is usually called the *likelihood* (written also as $p_\theta(x)$).
- $\pi(\theta)$ is the *prior*.
- $\pi(\theta|x)$ is the *posterior* (written also as $\pi_x(\theta)$).
- $p(x)$ is the *marginal* distribution of data.

The construction of a Bayesian model using densities is slightly less general than that in Theorems 2.1.1 and 2.1.2 but much more elementary. We just put $P_\theta(\mathrm{d}x) = p_\theta(\mathrm{d}x)$, $\Pi(\mathrm{d}\theta) = \pi(\theta)\mathrm{d}\theta$, $\Pi_x(\mathrm{d}\theta) = \pi_x(\theta)\mathrm{d}\theta$ and $P(\mathrm{d}x) = p(\mathrm{d}x)$. In fact the extra generality in Theorem 2.1.2 will be needed only when discussing sufficient ststistics in Subection 2.2.

Let us end our introduction with the following important remark. In (2.1.3), the marginal distribution $p(x)$ in the denominator is nothing but a norming constant of the posterior density. In Bayesian statistics the data (observation) x is considered as fixed. Consequently, Bayesians usually rewrite the Bayes formula as

$$\pi(\theta|x) \propto p(x|\theta)\pi(\theta), \tag{2.1.4}$$

where symbol '$\propto$' indicates that both sides are proportional as functions of θ. It turns out that there are techniques which allow to examine (recognize, compute or sample from) a probability density without knowning the norming constant.

2.2 Conditional independence and prediction

In this section we present a Bayesian decision-theoretic model of prediction. Let us first explain the methodological motivation. If scientific knowledge should be experimentally verifiable, then this rule applies also to statistical science. The results of statistical inference based on data X_1 have to be confronted with new data X_2. Thus

- X_1 is a *learning sample* (or training sample),
- X_2 is a *testing sample*.

We are to predict X_2 using the knowledge of X_1 and our statistical model. If the result of prediction is good, then the model is acceptable.

Adopting the Bayesian view, we assume that X_1 and X_2 are random variables with a joint probability distribution. To quantify the "goodness of prediction" we will use a loss function. The general scheme described in Section 1.1 has to be extended. The decision $a \in \mathcal{A}$ is chosen using the knowledge $X_1 = x_1$, so we consider the loss

$$\rho(\delta(x_1), X_2).$$

Function $\delta : \mathcal{X}_1 \to \mathcal{A}$ is called a *decision rule*. The loss function is $\rho : \mathcal{A} \to \mathcal{X}_2$. Typically we choose $\mathcal{A} = \mathcal{X}_2$ and $\rho(\delta(x_2), x_2)$ is some 'distance' or 'discrepancy' between the prediction $\delta(x_1)$ and the value to be predicted, x_2.

Conditional independence

As in Section 2.1, consider random variables ϑ and X with values in Polish spaces Θ and $\mathcal{X}$ (without loss of generality we can assume that they are defined on the canonical probability space $\Omega = \Theta \times \mathcal{X}$). Assume that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ and $X = (X_1, X_2)$. Random variables X_1 and X_2 are *conditionally independent* given ϑ iff for all Borel sets $B_1 \subseteq \mathcal{X}_1$ and $B_2 \subseteq \mathcal{X}_2$,

$$\mathbb{P}(X_1 \in B_1, X_2 \in B_2|\vartheta) = \mathbb{P}(X_1 \in B_1|\vartheta)\mathbb{P}(X_2 \in B_2|\vartheta) \tag{2.2.1}$$

almost surely. Equivalently, for every Borel $B_2 \subseteq \mathcal{X}_2$,

$$\mathbb{P}(X_2 \in B_2|\vartheta, X_1) = \mathbb{P}(X_2 \in B_2|\vartheta) \tag{2.2.2}$$

almost surely. The conditional independence is shortly denoted $X_1 \perp\!\!\!\perp X_2|\vartheta$.

We always assume that $\mathcal{X}_i$ and Θ are Polish spaces. Conditional independence can be concisely expressed in terms of the transition kernels. We have $X_1 \perp\!\!\!\perp X_2|\vartheta$ iff the joint probability distribution of (ϑ, X_1, X_2) is of the form

$$\mathbb{P}(\mathrm{d}\theta, \mathrm{d}x_1, \mathrm{d}x_2) = P_\theta^{(1)}(\mathrm{d}x_1)P_\theta^{(2)}(\mathrm{d}x_2)\Pi(\mathrm{d}\theta)$$

for some transition kernels $P^{(i)}$ from Θ to $\mathcal{X}_i$ such that $P_\theta^{(i)}(\mathrm{d}x_i) = \mathbb{P}(X_i \in \mathrm{d}x_i | \vartheta = \theta)$. In what follows, we will drop superscripts and write $P_\theta(\mathrm{d}x_1, \mathrm{d}x_2) = P_\theta(\mathrm{d}x_1)P_\theta(\mathrm{d}x_2)$.

Bayesian risk and posterior risk

Assume that $X_1 \perp\!\!\!\perp X_2 | \vartheta$ and consider the loss function $\rho(\delta(x_1), x_2)$. The *Bayesian risk* of the decision rule δ is defined as

$$\begin{aligned} r(\delta) = \mathbb{E}\rho(\delta(X_1), X_2) &= \iint_{\mathcal{X}_1 \times \mathcal{X}_2} \rho(\delta(x_1), x_2) P(\mathrm{d}x_1, \mathrm{d}x_2) \\ &= \iiint_{\Theta \times \mathcal{X}_1 \times \mathcal{X}_2} \rho(\delta(x_1), x_2) P_\theta(\mathrm{d}x_1) P_\theta(\mathrm{d}x_2) \Pi(\mathrm{d}\theta). \end{aligned} \tag{2.2.3}$$

If we put $\ell(a, \theta) = \mathbb{E}(\rho(a, X_2) | \vartheta = \theta) = \int_{\mathcal{X}_2} \rho(a, x_2) P_\theta(\mathrm{d}x_2)$ then

$$r(\delta) = \mathbb{E}\ell(\delta(X_1), \vartheta) = \iint_{\Theta \times \mathcal{X}_1} \ell(\delta(x_1), \theta) P_\theta(\mathrm{d}x_1) \Pi(\mathrm{d}\theta). \tag{2.2.4}$$

Interchanging the order of integration (or equvalently – the order of conditioning) we obtain the following expression:

$$\begin{aligned} r(\delta) = \mathbb{E}\mathbb{E}\left(\ell(\delta(X_1), \vartheta) | X_1\right) &= \int_{\mathcal{X}_1} \int_\Theta \ell(\delta(x_1), \theta) \Pi_{x_1}(\mathrm{d}\theta) P(\mathrm{d}x_1) \\ &= \int_{\mathcal{X}_1} r_{x_1}(\delta(x_1)) P(\mathrm{d}x_1), \end{aligned}$$

where $r_{x_1}(a) = \int_\Theta \ell(a, \theta) \Pi_{x_1}(\mathrm{d}\theta)$ is the *posterior risk* corresponding to action $a \in \mathcal{A}$.

If decision rule δ^* minimizes the Bayesian risk, that is $r(\delta^*) = \min_\delta r(\delta)$, then δ^* is called the *Bayes decision rule*.

2.2.5 Theorem. *If decision rule δ^* minimizes the posterior risk for all x_1, that is $r_{x_1}(\delta^*(x_1)) = \min_{a \in \mathcal{A}} r_{x_1}(a)$, then δ^* is the Bayes decision rule.*

Proof. For every δ we have pointwise inequality $r_{x_1}(\delta^*(x_1)) \leq r_{x_1}(\delta(x_1))$, so inequality between the integrals follows, $\int r_{x_1}(\delta^*(x_1)) P(\mathrm{d}x_1) \leq \int r_{x_1}(\delta(x_1)) P(\mathrm{d}x_1)$. □

Theorem 2.2.5 is as obvious as it is important. It says that conditioning on x_1 reduces the problem of finding an optimal decision rule to the easier problem of minimization over the space of decisions.

2.2.6 REMARK. In our presentation we underline the role of prediction and testing sample. The posterior risk can be expressed either as an expectation of loss ρ with respect to the *predictive* distribution,

$$r_{x_1}(a) = \mathbb{E}(\rho(a, X_2)|X_1 = x_1),$$

or as an expectation of loss ℓ ect to the *posterior* distribution

$$r_{x_1}(a) = \mathbb{E}(\ell(a, \vartheta)|X_1 = x_1).$$

If ℓ is related to ρ via $\ell(a, \theta) = \mathbb{E}(\rho(a, X_2)|\vartheta = \theta)$, both formulas are equivalent. However, the standard approach is to begin with a function $\ell : \mathcal{A} \times \Theta$ and define the Bayesian risk by (2.2.4), without considering X_2 and ρ. The term 'loss function' is then reserved for ℓ. We will use symbol $\rho(a, \mathcal{X}_2)$ to denote loss which depends on action a and future observation X_2. Loss which depends on action a and parameter θ will be denoted $\ell(a, \theta)$.

2.2.7 REMARK. Our approach is Bayesian. From the frequentist viewpoint, the Bayesian risk is obtained in a different way, by taking integrals in a reverse order:

$$\begin{aligned} r(\delta) = \mathbb{E}\mathbb{E}\left[\ell_\vartheta(\delta(X_1))|\vartheta\right] &= \int_\Theta \int_{\mathcal{X}_1} \ell(\delta(x_1), \theta) P_\theta(\mathrm{d}x_1)\Pi(\mathrm{d}\theta) \\ &= \int_\Theta R_\theta(\delta)\Pi(\mathrm{d}\theta), \end{aligned}$$

where $R_\theta(\delta) = \int_{\mathcal{X}_1} \ell(\delta(x_1), \theta)P_\theta(\mathrm{d}x_1)$ is the frequentist ('Berkeley') risk function, defined with no reference to Π. The prior Π is then viewed as a 'weighting' distribution.

Sufficiency

From a Bayesian perspective, sufficiency is nothing but a special case of conditional independence. Let us describe details, beginning with the 'usual' (frequentist) characterization of sufficient statistics.

We consider only a learning sample, renamed to X, with values in $\mathcal{X}$. We have a family of probability distributions $\{P_\theta(\mathrm{d}x) : \theta \in \Theta\}$. A *statistic* is a (measurable) function $T : \mathcal{X} \to \mathcal{T}$. If the probability distributions P_θ have densities p_θ with respect to a common measure, $P_\theta(\mathrm{d}x) = p_\theta(x)\mathrm{d}x$ then the discussion simplifies. We can use the well-known factorization criterion. Statistic $T : \mathcal{X} \to \mathcal{T}$ is sufficient iff there are versions of densities which satisfy

$$p_\theta(x) = f_\theta(T(x))g(x)$$

for some functions $f : \Theta \times \mathcal{T} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$. The Bayes formula gives

$$\pi_x(\theta) \propto f_\theta(T(x))g(x)\pi(\theta) \propto f_\theta(T(x))\pi(\theta).$$

The posterior depends on x only through $T(x)$. This implies that $X \perp\!\!\!\perp \vartheta|T(X)$.

2.2.8 Proposition. *If statistic T is sufficient then $X \perp\!\!\!\perp \vartheta | T$.*

Proposition 2.2.8 gives a direct and intuitive interpretation of sufficiency. For a Bayesian, posterior distribution contains all available information about θ. If we have $\mathbb{P}(\mathrm{d}\theta|X) = \mathbb{P}(\mathrm{d}\theta|T(X))$ then the knowledge of $T(x) = t$ is as good as the knowledge of x itself.

The converse of Proposition 2.2.8 is "almost" true.

2.2.9 Proposition. *If $X \perp\!\!\!\perp \vartheta | T$ then there is a subset $\tilde{\Theta}$ of Θ with $\Pi(\tilde{\Theta}) = 1$ such that T is sufficient in the 'usual' (frequentist) sense with respect to the family $\{P_\theta : \theta \in \tilde{\Theta}\}$.*

We omit the proof. The restriction to an 'almost sure' subset $\tilde{\Theta}$ is necessary, because conditional independence is a property of the joint distribution of (ϑ, X, T), and thus ignores the behaviour of P_θ on null subsets of Θ.

The role of sufficiency for statistical decisions is explained by the following fact.

2.2.10 Proposition (Sufficiency and decisions)**.** *Assume that statistic T is sufficient. For an arbitrary loss function ℓ, the Bayes decision rule δ^* depends on x only through $T(x)$.*

Proof. This is obvious, since the Bayesian risk $R(\delta)$ is an integral with respect to the posterior $\mathbb{P}(\mathrm{d}\theta|X) = \mathbb{P}(\mathrm{d}\theta|T(X))$ so it depends on X only through $T(X)$. □

In the frequentist statistics, things get a little bit more complicated. We need an additional assumption of convexity.

2.2.11 Proposition (Blackwell-Rao)**.** *Assume that statistic T is sufficient. Assume that the space of actions $\mathcal{A}$ is a convex subset of $\mathbb{R}^d$ and $\ell(a, \theta)$ is a convex function of a for every θ. For a decision rule δ, define*

$$\delta'(t) = \mathbb{E}(\delta(X)|T = t).$$

Then δ' is a decision rule with the frequentist risk uniformly less than that of δ: $R_\theta(\delta') \leq R_\theta(\delta)$ for all θ.

Proof. By Jensen inequalty,

$$\begin{aligned} R_\theta(\delta') = \mathbb{E}\left(\ell(\delta'(T), \theta)\middle|\theta\right) &= \mathbb{E}\left(\ell(\mathbb{E}(\delta(X)|T), \theta)\middle|\theta\right) \\ &= \mathbb{E}\left(\ell(\mathbb{E}(\delta(X)|T, \theta), \theta)\middle|\theta\right) \\ &\leq \mathbb{E}\left(\mathbb{E}(\ell(\delta(X), \theta)|T, \theta)\middle|\theta\right) \\ &\leq \mathbb{E}\ell(\delta(X), \theta) = R_\theta(\delta). \end{aligned}$$

In the second line, we used sufficiency of T. □

* Sufficiency revisited

In this subsection we present a more rigorous treatment of sufficiency, based on regular versions of conditional probability distributions. Our starting point is the standard frequentist definition of sufficiency, and not the factorization theorem.

As usual in these notes, we assume that $\mathcal{X}$ and $\mathcal{T}$ are Polish spaces equipped with Borel σ-fields. Statistic T is sufficient if, for every (Borel) $B \subseteq \mathcal{X}$, there exists a (meaurable) function $\mathcal{T} \ni t \mapsto P_t(B)$ which, for every θ, is a version of the conditional probability distribution $P_\theta(B|T=t)$ (the point is that $P_t(B)$ is the same for all θ). Statistic T transports measure P_θ on $\mathcal{X}$ to measure Q_θ on $\mathcal{T}$ according to the formula $Q_\theta(A) = P_\theta(T^{-1}(A))$ for any Borel set $A \subseteq \mathcal{T}$. By definition of conditional probability, statistic T is sufficient iff there is a function $t \mapsto P_t(B)$ such that

$$\int_A P_t(B) Q_\theta(\mathrm{d}t) = P_\theta(B \cap T^{-1}(A))$$

for all θ and $A \subseteq \mathcal{T}$. It may be shown that $P_t(B)$ can be chosen to be a *regular* version of conditional probability, that is a transition kernel (because $\mathcal{X}$ is Polish). Now we proceed to a Bayesian model and introduce a prior Π on Θ. The joint probability distribution of $(\vartheta, X, T=T(X))$ is a measure on $\Theta \times \mathcal{X} \times \mathcal{T}$ given by

$$\mathbb{P}(C \times B \times A) = \int_C P_\theta(B \cap T^{-1}(A)) \Pi(\mathrm{d}\theta).$$

If T is sufficient then we can rewrite this equation as

$$\mathbb{P}(C \times B \times A) = \int_C \int_A P_t(B) Q_\theta(\mathrm{d}t) \Pi(\mathrm{d}\theta).$$

Theorem 2.1.2 ensures that there exists a kernel $\Pi_t(\mathrm{d}\theta)$ such that $Q_\theta(\mathrm{d}t)\Pi(\mathrm{d}\theta) = \Pi_t(\mathrm{d}\theta)Q(\mathrm{d}t)$, where $Q(\mathrm{d}t)$ is the marginal distribution of $T(X)$. We obtain

$$\mathbb{P}(C \times B \times A) = \int_A \int_C P_t(B) \Pi_t(\mathrm{d}\theta) Q(\mathrm{d}t) = \int_A P_t(B) \Pi_t(C) Q(\mathrm{d}t).$$

This means that the transition kernel $\mathcal{T} \to \mathcal{X} \times \Theta$ is a product measure for every t. We have shown that $X \perp\!\!\!\perp X|T$, that is proved Proposition 2.2.8.

In view of (2.2.1), $X \perp\!\!\!\perp \vartheta|T$ is equivalent to $\mathbb{P}(\mathrm{d}\theta|X) = \mathbb{P}(\mathrm{d}\theta|T)$. Put differently, if $\Pi_t(\mathrm{d}\theta)$ is a transition kernel $\mathcal{T} \to \Theta$ such that $\Pi_t(C) = \mathbb{P}(\vartheta \in C|T=t)$ then $\Pi_{T(x)}(C)$ is a version of $\mathbb{P}(\vartheta \in C|X=x)$. Indeed, we have $\mathbb{1}(T(x) \in \mathrm{d}t)P(\mathrm{d}x) = P_t(\mathrm{d}x)Q(\mathrm{d}t)$ by Theorem 2.1.2. Therefore

$$\begin{aligned}
\int_B \Pi_{T(x)}(C) P(\mathrm{d}x) &= \int_B \int_{\mathcal{T}} \Pi_t(C) \mathbb{1}(T(x) \in \mathrm{d}t) P(\mathrm{d}x) \\
&= \int_{\mathcal{T}} \int_B \Pi_t(C) P_t(\mathrm{d}x) Q(\mathrm{d}t) = \int_{\mathcal{T}} \Pi_t(C) P_t(B) Q(\mathrm{d}t) \\
&= \mathbb{P}(C \times B \times \mathcal{T}) = \mathbb{P}(\vartheta \in C, X \in B),
\end{aligned}$$

and we have obtained the condition which appears in the definition of conditional probability.

Chapter 3

Examples of Bayesian models

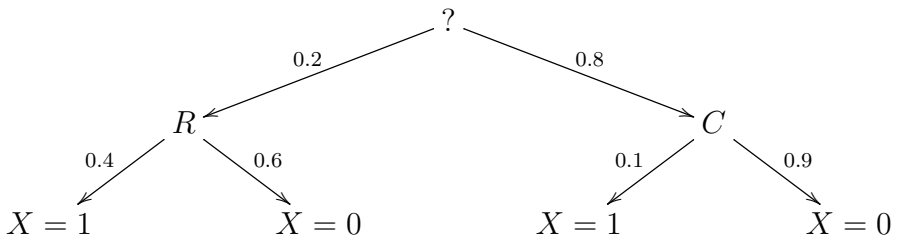
3.1 Introductory example

Let us begin with a extremely simplified example which well illustrates the roles played by the main characters in the Bayesian world: the likelihood, the prior, the posterior, the predictive.

3.1.1 EXAMPLE. The insurer signs a contract with a client driver. There are two types of drivers: *R*isky and *C*autious. In a coming year, a *R*isky will cause an accident and thus incure a loss with probability 0.4. For a *C*autious, the probability of loss is 0.1. The insurer does not know whether the client is of the *R*isky type or is *C*autious. However, she knows that (in the population of her potential clients) there is 1 *R*isky for 4 *C*autious'. The situation corresponds to the following Bayesian model. Let $X = 1$ stand for a 'loss' and $X = 0$ for 'no loss' event.

- The **likelihood**:
$$\mathbb{P}(X = 1|C) = 0.1, \qquad \mathbb{P}(X = 1|R) = 0.4.$$
- The **prior**:
$$\mathbb{P}(C) = 0.80, \qquad \mathbb{P}(R) = 0.20.$$
- The **marginal** computed via the rule of total probability:
$$\begin{aligned}\mathbb{P}(X = 1) &= \mathbb{P}(X = 1|C)\mathbb{P}(C) + \mathbb{P}(X = 1|R)\mathbb{P}(R)\\ &= 0.1 \times 0.80 + 0.4 \times 0.20 = 0.16.\end{aligned}$$
- The **posterior** computed via the Bayes rule:
$$\mathbb{P}(R|X = 1) = \frac{\mathbb{P}(X = 1|R)\mathbb{P}(R)}{\mathbb{P}(X = 1)} = 0.5.$$

The following tree diagram describes the probability distribution.



Note that this is a 2 stage random experiment, in which the insurer does not know the outcome of the 1st stage (R or C). After a year she observes the outcome of the 2nd stage: $X = 0$ – 'loss' or $X = 0$ – 'no loss'.

To predict future losses we have to make additional assumptions. Let random variables X_1 and X_2 correspondsto losses in two years for the same client. Assume that 'the type of the client is unchanged and his style of driving in the second year is independent of what has happened in the first year. Put differently,

$$\mathbb{P}(X_2 = 1|C, X_1) = 0.1, \qquad \mathbb{P}(X = 1|R, X_1) = 0.4$$

(variables X_2 is conditionally independent of X_1 and has the same probability distribution). Then we can make a reasonable prediction.

- The **predictive** distribution:

$$\begin{aligned}\mathbb{P}(X_2 = 1|X_1 = 1) &= \mathbb{P}(X_2 = 1|C)\mathbb{P}(C|X_1 = 1) + \mathbb{P}(X_2 = 1|R)\mathbb{P}(R|X_1 = 1)\\ &= 0.1 \times 0.5 + 0.4 \times 0.5 = 0.25.\end{aligned}$$

In the standard Bayesian notation, the 1st stage of the random experiment is described as drawing at random a parameter (say ϑ with values $\theta \in \{C, R\}$), which governs the 2nd stage.

Let us underline that in this example the prior has an objective interpretation, which is not the orthodox Bayesian standpoint. △

3.2 Typical conjugate models

In many typical Bayesian models, the *prior* probability distribution is chosen in such a way that the *posterior* is easy to compute. This is usually achieved if the likelihood (probability distribution of observed variables) and the prior belong to the so-called *conjugate families* of distributions. We defer the general definition and begin with a few most popular examples.

In these examples the denominator $p(x)$ in the Bayes formula is relatively easy, but it is not always necessary to compute it. If we are interested only in the posterior, then we can omit factors independent of θ (even if they depend on x) and rewrite Bayesa formula in a simplified manner:

$$\pi_x(\theta) \propto \pi(\theta) p_\theta(x).$$

Symbol $\propto$ means that, LHS and RHS are proportional as functions of θ.

3.2.1 EXAMPLE (Binomial likelihood and beta prior). Assume that we observe outcomes of n Bernoulli trials with unknown probability of success θ. Probability distribution of binary (0-1) variates $X_1, \ldots, X_n$ is $P_\theta(X_1 = x_1, \ldots, X_n = x_n) = p_\theta(x_1, \ldots, x_n) = \theta^s(1-\theta)^{n-s}$, where $s = \sum x_i$. The number of successes $S = \sum X_i$ is a sufficient statistic and has the binomial distribution $\mathrm{Bin}(n, \theta)$. Assume that the prior distribution is beta, $\theta \sim \mathrm{Beta}(\alpha, \beta)$, that is

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad (0 < \theta < 1).$$

The posterior density is

$$\pi_{x_1,\ldots x_n}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^s(1-\theta)^{n-s} = \theta^{\alpha+s-1}(1-\theta)^{\beta+n-s-1}.$$

Thus the posterior distribution is also beta, $\theta|s \sim \mathrm{Beta}(\alpha + s, \beta + n - s)$. △

Figure 3.1 shows *prior* and posterior distributions in Example 3.2.1 for several sample sizes n. The prior is the uniform distribution $\mathrm{U}(0, 1)$, which is a special case of beta, $\mathrm{Beta}(1, 1)$.

3.2.2 EXAMPLE (Poisson likelihood and gamma prior). Assume that $X_1, \ldots, X_n$ are independent with Poisson distribution $\mathrm{Poiss}(\theta)$. Let the prior for θ be $\mathrm{Gamma}(\alpha, \lambda)$, that is

$$\pi(\theta) \propto \theta^{\alpha-1}\mathrm{e}^{-\lambda\theta}, \qquad (\theta > 0).$$

The posterior density is

$$\pi_{x_1,\ldots x_n}(\theta) \propto \theta^{\alpha-1}\mathrm{e}^{-\lambda\theta}\mathrm{e}^{-n\theta}\theta^s = \theta^{\alpha+s-1}\mathrm{e}^{-(\lambda+n)\theta},$$

where $s = \sum x_i$. Therefore the posterior is $\mathrm{Gamma}(\alpha + s, \lambda + n)$ and depends on X_is only through values of the sufficient statistic $S = \sum X_i$. △

This model is used in insurance mathematics and the so-called 'credibility theory'. Interpretation is similar as in the simplified Example 3.1.1, but the Poisson/Gamma model is more flexible and more realistic. Assume that X_i denotes the number of losses in ith year of insurance. Parameter θ is the mean number of losses per year. To simplify considerations, imagine that we analyse data for a single client of an insurance company. The parameter θ describes the 'level of risk' for this client. The prior distribution describes variability of the 'risk parameter' in the population of clients. Any new contract is regarded as a *two-stage*

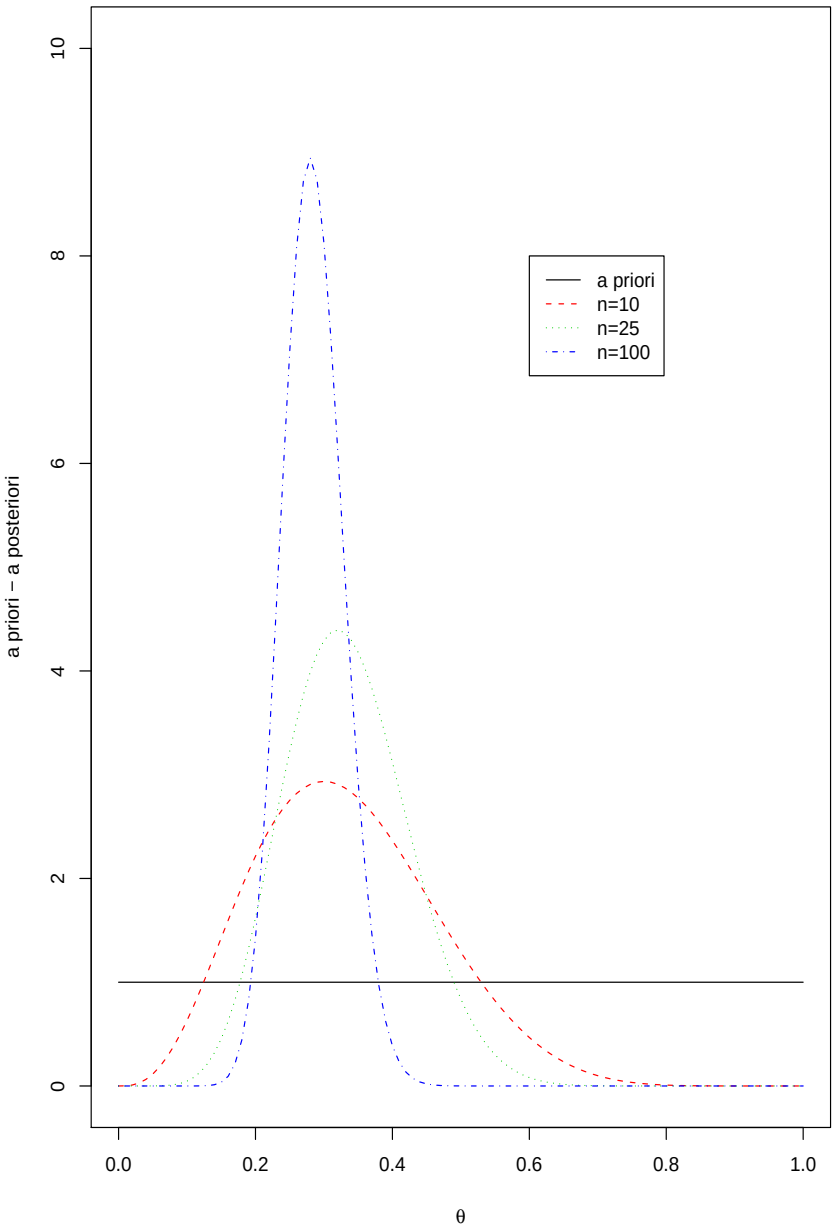


Figure 3.1: Prior and posterior distributions in Example 3.2.1.

random experiment. In the *first stage* the insurer signs a contract with a client 'drawn at random from a pool of potential clients'. At this stage the risk random variable θ is sampled, but not revealed to the insurer. In the *second stage* θ is kept fixed and the insurer observes losses in subsequent years, that is variables $X_1, \ldots, X_i, \ldots$. In this way the insurer learns more and more about the risk parameter θ. The state of knowledge (or lack of knowledge) about θ is described by the posterior distribution.

3.2.3 EXAMPLE (Normal likelihood and normal prior for the mean). Let the observed variables $X_1, \ldots, X_n$ be sampled from the normal distribution $\mathrm{N}(\mu, \sigma^2)$. Assume that σ^2 is known, and unknown parameter μ has normal prior $\mathrm{N}(m, v^2)$, that is

$$\pi(\mu) \propto \exp\left[-\frac{1}{2v^2}(\mu - m)^2\right].$$

The posterior density is

$$\pi_{x_1,\ldots,x_n}(\mu) \propto \exp\left[-\frac{1}{2v^2}(\mu - m)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right],$$

which can be expressed as follows:

$$\pi_{x_1,\ldots,x_n}(\mu) \propto \exp\left[-\frac{nv^2 + \sigma^2}{2\sigma^2 v^2}\Big(\mu - \frac{nv^2\bar{x} + \sigma^2 m}{nv^2 + \sigma^2}\Big)^2\right].$$

Of course, $\bar{x}$ denotes $\sum x_i/n$. The computation in this example is based on the 'canonical form of a quadratic function'. Note that the intercept in the exponent is absorbed into the symbol $\propto$. We have shown that the posterior is a normal distribution,

$$\mathrm{N}\Big(\frac{nv^2\bar{x} + \sigma^2 m}{nv^2 + \sigma^2}, \frac{\sigma^2 v^2}{nv^2 + \sigma^2}\Big).$$

△

3.2.4 EXAMPLE (Normal likelihood and inverse gamma prior for the variance). Assume $X_1, \ldots, X_n \sim_{\text{i.i.d.}} \mathrm{N}(\mu, \sigma^2)$ where μ is known. For σ^2 choose the prior known as inverse gamma, $\sigma^2 \sim \mathrm{IG}(\alpha, \lambda)$. By definition, this means that $\sigma^{-2} \sim \mathrm{Gamma}(\alpha, \lambda)$. (The inverse variance $\nu = \sigma^{-2}$ is named *precision*.) Thus

$$\pi(\nu) \propto \nu^{\alpha-1}\exp\left(-\lambda\nu\right].$$

It is easy to check that the posterior is

$$\pi_{x_1,\ldots,x_n}(\nu) \propto \nu^{\alpha+n/2-1}\exp\left[-\left(\lambda + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)\nu\right].$$

Thus $\sigma^2|x \sim \mathrm{IG}(\alpha + n/2, \lambda + \sum(x_i - \mu)^2/2)$. △

3.2.5 EXAMPLE (Normal likelihood, joint prior for the mean and the variance). Consider the family of normal distributions $\mathrm{N}(\theta, \sigma^2)$ with both parameters unknown. Let the prior for (θ, σ^2) be described as follows. Marginal prior for σ^2 is inverse gamma, $\sigma^2 \sim \mathrm{IG}(\alpha, \lambda)$. Conditional (prior) distribution of θ is $\theta|\sigma^2 \sim \mathrm{N}(\mu, r\sigma^2)$ for some μ and $r > 0$. The posterior turns out to be of the same form, with different parameters replacing μ, α, λ and r. Note that the conjugate 2-dim prior is not a product distribution. See Problem 5. △

3.2.6 EXAMPLE (Multinomial likelihood, Dirichlet prior). The multivariate counterpart of Beta is the Dirichlet distribution. We say that d-dimensional random variable θ has the *Dirichlet distribution*,

$$\theta = (\theta_1, \ldots, \theta_d) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_d),$$

if $\theta_1 + \cdots + \theta_d = 1$ and $\theta_1, \ldots, \theta_{d-1}$ have joint probability density

$$p(\theta_1, \ldots, \theta_{d-1}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_d)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_d)} \theta_1^{\alpha_1 - 1} \ldots \theta_{d-1}^{\alpha_{d-1} - 1} (1 - \theta_1 - \ldots \theta_{d-1})^{\alpha_d - 1}.$$

Parameters $\alpha_1, \ldots, \alpha_r$ can be arbitrary nonnegative reals. Note that for $d = 2$, Dirichlet is just Beta:

$$\theta_1 \sim \mathrm{Beta}(\alpha_1, \alpha_2) \quad \text{iff} \quad (\theta_1, 1 - \theta_1) \sim \mathrm{Dir}(\alpha_1, \alpha_2).$$

A basic property of Dirichlet is the conjugacy with multinomial distribution. If $(N_1, \ldots, N_d|\theta) \sim \mathrm{Mult}(\theta_1, \ldots, \theta_d)$ and $\theta \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_d)$ then $(\theta|N_1, \ldots, N_d) \sim \mathrm{Dir}(\alpha_1 + N_1, \ldots, \alpha_d + N_d)$.

△

In the above examples we were interested solely in the posterior distribution and paid no attention to the marginal (the denominator in the Bayes formula). However, sometimes the marginal is of independent interest. In conjugate models, the predictive distribution has the same form as the marginal. Below we compute the marginal in some previously considered examples.

3.2.7 EXAMPLE (Bernoulli likelihood and beta prior). Recall Example 3.2.1. If X is a binary variable with $P_\theta(X = x) = p_\theta(x) = \theta^x(1 - \theta)^{1-x}$ and the prior distribution is beta, $\theta \sim \mathrm{Beta}(\alpha, \beta)$, then marginally $\mathbb{P}(X = 1) = \alpha/(\alpha + \beta)$ and $\mathbb{P}(X = 0) = \beta/(\alpha + \beta)$. △

The Binomial/Beta model will be revisited in Example 6.2.3, where the joint marginal is interpeted as an urn scheme.

3.2.8 EXAMPLE (Poisson likelihood and gamma prior). Let X has the Poisson distribution $\mathrm{Poiss}(\theta)$ and the prior for θ be $\mathrm{Gamma}(\alpha, \lambda)$. Recall Example 3.2.2 and the interpretation of the model. In insurance applications, X is the number of losses in a year. The marginal distribution of X corresponds to the number of losses for a single client selected at random

from the population (described by the gamma prior).

$$\begin{aligned} f(x) = \mathbb{P}(X = x) &= \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\theta\lambda} \cdot \frac{\theta^x}{x!} e^{-\theta} \mathrm{d}\theta \\ &= \frac{\lambda^\alpha}{x!\Gamma(\alpha)} \int_0^\infty \theta^{x+\alpha-1} e^{-(\lambda+1)\theta} \mathrm{d}\theta \\ &= \frac{\lambda^\alpha}{x!\Gamma(\alpha)} \cdot \frac{\Gamma(x+\alpha)}{(\lambda+1)^{x+\alpha}} \\ &= \frac{(x+\alpha-1)(x+\alpha-2)\cdots(\alpha+1)\alpha}{x!} \left(\frac{\lambda}{\lambda+1}\right)^\alpha \left(\frac{1}{\lambda+1}\right)^x \\ &= \binom{x+\alpha-1}{x} \Big(\frac{\lambda}{\lambda+1}\Big)^\alpha \Big(\frac{1}{\lambda+1}\Big)^x. \end{aligned}$$

Let $\lambda/(\lambda+1) = p$ and $1/(\lambda+1) = 1-p$. Then

$$\mathbb{P}(X = x) = \binom{-\alpha}{x} p^\alpha (p-1)^x. \tag{3.2.9}$$

This is a *Negative Binomial* distribution, $\mathrm{Bin}^-(\alpha, p)$. △

3.2.10 EXAMPLE (Normal likelihood and normal prior for the mean). Let the observed variable X have the normal distribution $\mathrm{N}(\mu, \sigma^2)$ and the unknown parameter μ have normal prior $\mathrm{N}(m, v^2)$. The marginal for X is $\mathrm{N}(m, v^2 + \sigma^2)$. △

In conjugate models the predictive distribution can be immediately derived from the formulas for the posterior and the marginal.

3.3 Conjugate priors

If the likelihood belongs to an exponential family of distributions, then there exists a special family of priors exceptionally friendly from the computational point of view. Recall that a density p_θ belongs to an *exponential family* if it has the form

$$p_\theta(x) = \exp\left\{ \sum_{j=1}^k T_j(x) g_j(\theta) + g_0(\theta) \right\} h(x) = \exp\left\{ T(x)^\top g(\theta) + g_0(\theta) \right\} h(x),$$

where $g(\theta) = (g_1(\theta), \ldots, g_k(\theta))^\top$ and $T(x) = (T_1(x), \ldots, T_k(x))^\top$. The important property of such distributions is the existence of sufficient statistics of fixed dimension k, independenent of the sample size. Indeed, the likelihood for $X_1, \ldots, X_n \sim_{\text{i.i.d.}} p_\theta$ is

$$\prod_{i=1}^n p_\theta(x_i) = \exp\left\{ \sum_{j=1}^k \left(\sum_{i=1}^n T_j(x_i) \right) g_j(\theta) + n g_0(\theta) \right\} \prod_{i=1}^n h(x_i)$$

The factorization criterion shows that $\sum_{i=1}^n T(x_i)$ is a k-dimensional sufficient statistic.

We define the family of *conjugate prior* densities by

$$\begin{aligned}\pi(\theta) = \pi_{t_0,n_0}(\theta) &= c(t_0,n_0)\exp\left\{\sum_{j=1}^k t_{0j}g_j(\theta) + n_0 g_0(\theta)\right\}\\ &= c(t_0,n_0)\exp\Big\{t_0^\top g(\theta) + n_0 g_0(\theta)\Big\},\end{aligned}$$

provided that the integral of this function over Θ is finite (then there exists a normalizing constant $c(t_0,n_0)$). The key feature of a conjugate families is that the posterior has the same form as the prior, but with different parameters. Indeed,

$$\begin{aligned}\pi(\theta|x_1,\ldots,x_n) &\propto \prod_{i=1}^n p_\theta(x_i)\pi(\theta) \propto \exp\left\{\sum_{i=1}^n T(x_i)^\top g(\theta) + n g_0(\theta)\right\}\exp\left\{t_0^\top g(\theta) + n_0 g_0(\theta)\right\}\\ &= \exp\left\{\sum_{i=1}^n \Big(T(x_i) + t_0\Big)^\top g(\theta) + (n+n_0)g_0(\theta)\right\}.\end{aligned}$$

Thus π_{t_0,n_0} is updated to $\pi_{t_0+n\bar{t},n_0+n}$, where $n\bar{t} = \sum_{i=1}^n T(x_i)$. It is interesting to note that t_0 and n_0 play the roles analogous to $\sum T(x_i)$ and n, respectively. One can imagine that t_0 is a sum of values of T for n_0 "virtual observations". This is a convenient way of translating the prior knowledge of an expert into the prior distribution.

In the sequel we will use the *natural parametrization* of an exponential family, i.e. assume (w.l.o.g.) that $\Theta \subseteq \mathbb{R}^k$, $g_j(\theta) = \theta_j$ and $g_0(\theta) = -\psi(\theta)$. The formula for p_θ simplifies to

$$\begin{aligned}p_\theta(x) &= \exp\left\{\sum_{j=1}^k T_j(x)\theta_j - \psi(\theta)\right\}h(x),\\ \psi(\theta) &= \log\int_{\mathcal{X}} \exp\left\{\sum_{j=1}^k T_j(x)\theta_j\right\}h(x)\mathrm{d}x\end{aligned}$$

It is well-known that $\psi(\theta)$ is the comulant generating function and, in particular, $\mathbb{E}(T(X)|\theta) = \nabla\psi(\theta)$ and $\mathrm{VAR}(T(X)|\theta) = \nabla\nabla^\top\psi(\theta)$. If we put $\mu(\theta) = \mathbb{E}(T(X)|\theta)$ then

$$\begin{aligned}\mathbb{E}\mu(\theta) &= \int_\Theta \mu(\theta)\pi_{t_0,n_0}(\theta)\mathrm{d}\theta = \frac{t_0}{n_0},\\ \mathbb{E}[\mu(\theta)|x_1,\ldots,x_n] &= \int_\Theta \mu(\theta)\pi_{t_0+n\bar{t},n_0+n}(\theta)\mathrm{d}\theta = \frac{t_0+n\bar{t}}{n_0+n},\end{aligned} \tag{3.3.1}$$

provided that Θ is an open covex subset of $\mathbb{R}^k$. Of course, $\mathbb{E}\mu(\theta) = \mathbb{E}T(X)$ (expectation w.r.t. $p_\theta(x)\pi(\theta)$.

Sketch of proof of Equation (3.3.1). Assume that $k = 1$ and $\Theta = \mathbb{R}$. Then $\psi'(\theta) = \mu(\theta)$ and $\pi'(\theta) = (t_0 - n_0\mu(\theta))\pi(\theta)$. It can be shown (and it is rather intuitively clear) that $\int \pi'(\theta)\mathrm{d}\theta = 0$, so $t_0 = n_0 \int \mu(\theta)\pi(\theta)\mathrm{d}\theta$. The first equation in (3.3.1) follows and the second one is an immediate consequence of the first. □

All models considered in Section 3.2 are examples of cojugate priors. There are a few examples of likelihoods which do not belong to any exponential family, but have well-behaved priors similar to conjugate priors.

3.3.2 EXAMPLE (Uniform likelihood and Pareto prior). Let $X_1, \ldots, X_n|\theta \sim_{\text{i.i.d.}} \mathrm{U}(0, \theta)$ and for the prior choose the Pareto distribution with the density $\pi(\theta) \propto \theta^{-n_0}\mathbb{1}(\theta > t_0)$. Since $p_\theta(x_1, \ldots, x_n) = \theta^{-n}\mathbb{1}(\max(x_1, \ldots, x_n) < \theta)$, the posterior is the Pareto with the density $\propto \theta^{-n_0-n}\mathbb{1}(\theta > t_0 \vee \max(x_1, \ldots, x_n))$. Although the family of uniforms is not an exponential family, there does exist a sufficient statistic of fixed dimension 1, namely $T = \max(x_1, \ldots, x_n)$. The construction of quasi-conjugate prior is based on this statistic. △

Conjugate priors have computational advantages. The obvious disadvantage is a narrow choice of conjugate models. A possible remedy is to use mixtures of conjugate priors.

3.3.3 EXAMPLE (Mixtures of conjugate priors). Consider an exponential family (with natural parametrization)

$$p_\theta(x) = \exp\Big\{T(x)^\top\theta - \psi(\theta)\Big\}h(x)$$

and the family of conjugate priors

$$\pi_{t_0,n_0}(\theta) = c(t_0, n_0)\exp\Big\{t_0^\top\theta - n_0\psi(\theta)\Big\}.$$

If we choose a mixture of such distributions as a prior, i.e. let

$$\bar{\pi}(\theta) = \sum_{j=1}^{k} w_j \pi_{t_{0j},n_{0j}}(\theta),$$

for some choice of hyper-parameters (t_{0j}, n_{0j}) and weights $w_j > 0$ $(\sum w_j = 1)$ then the posterior is also a mixture,

$$\bar{\pi}_{x_1,\ldots,x_n}(\theta) = \sum_{j=1}^{k} w'_j \pi_{t_{0j}+n\bar{t},n_{0j}+n}(\theta),$$

where the posterior weights are $w'_j \propto w_j c(t_{0j}, n_{0j})/c(t_{0j} + n\bar{t}, n_{0j} + n)$. Verification of this fact is left as an exercise. △

3.4 Estimation and prediction

For a true Bayesian, the computation of the posterior distribution is the final result of statistical inference. However, from a more pragmatic viewpoint, the posterior is only a basis for choosing optimal decisions. Statistical problems can be classified according to the "type of decisions" and the appropriate loss functions. *Estimation* is the statistical equivalent of "approximation". In the Bayesian world it is always approximation of one random variable with a function of another random variable. However, we prefer to speak of *estimation* when approximating a hypothetical quantity such as a parameter of the model. If we approximate a future observation, we use the term *prediction*.

Let us begin with estimation. We are to approximate a function of unobserved θ with a function of observed x (data). In the Bayesian world, both $\theta \in \Theta$ and $x \in \mathcal{X}$ are sampled values of random variables ϑ and X.

Let $g : \Theta \to \mathbb{R}$ and $\hat{g} : \mathcal{X} \to \mathbb{R}$. For a space of actions we take $\mathbb{R}$ and we can use one of the loss functions defined in Section 1.1 to quantify "how well $\hat{g}(x)$ approximates $g(\theta)$". Decision rule $\hat{g}$ is called an estimator. The Bayes estimator, denoted by g^*, is the minimizer of the corresponding Bayes risk

$$r(\hat{g}) = \mathbb{E}\ell(\hat{g}(X), \vartheta) = \iint_{\Theta\times\mathcal{X}} \ell(\hat{g}(x), \theta) P_\theta(\mathrm{d}x)\Pi(\theta).$$

In the sequel, we drop the unnessary distinction between ϑ and θ.

3.4.1 EXAMPLE (Square loss). Most often used criterion is

$$\ell(a, \theta) = (g(\theta) - a)^2.$$

The Bayes estimator is the posterior mean, $g^*(x) = \mathbb{E}[g(\theta)|X = x]$. It follows immediately from Theorem 2.2.5 and Example 1.1.1. The risk corresponding to the square loss is called MSE (mean square error). △

3.4.2 EXAMPLE (Weighted square loss). A slight generalization is

$$\ell(a, \theta) = w(\theta)(g(\theta) - a)^2,$$

where w is a known weight function. The Bayes estimator is the weighted posterior mean, $g^*(x) = \dfrac{\mathbb{E}[w(\theta)g(\theta)|X = x]}{\mathbb{E}[w(\theta)|X = x]}$.

An interesting special case is relative square error, especially suitable in estimating positive quantity $g(\theta)$ of unknown order of magnitude.

$$\ell(a, \theta) = \left(\frac{g(\theta) - a}{g(\theta)}\right)^2.$$

△

3.4.3 EXAMPLE ("Precautionary" loss). This loss function is based on a similar idea as the relative error. The estimated values close to zero are heavily penalized.

$$\ell(a,\theta) = \left(\frac{g(\theta)-a}{a}\right)^2.$$

△

3.4.4 EXAMPLE (Quantile). Let

$$\ell(a,\theta) = a + \frac{g(\theta)-a}{1-p}\mathbb{1}(g(\theta) > a).$$

The Bayes estimator is the pth quantile of the posterior distribution, $\mathbb{P}(g(\theta) \leq g^*(x)|x) = p$ provided that $\mathbb{P}(g(\theta) = g^*(x)|x) = 0$. It follows from Theorem 2.2.5 and Example 1.1.2. △

3.4.5 EXAMPLE (LINEX). Let $\kappa > 0$ and

$$\ell(a,\theta) = \exp\{\kappa(g(\theta)-a)\} - \kappa(g(\theta)-a) - 1,$$

The corresponding Bayes estimator is $g^*(x) = \frac{1}{\kappa}\log\mathbb{E}\left(\exp\{\kappa g(\theta)\}\middle| X = x\right)$. △

Prediction

To fix ideas, consider the typical model with $X_1, \ldots, X_n, X_{n+1}$ conditionally i.i.d., given θ. We treat $X_{1:n} = (X_1, \ldots, X_n)$ as data and X_{n+1} as a future observation. We are to predict X_{n+1} after observing $X_{1:n} = x_{1:n}$. Note that θ is only introduced to "convey information" from $X_{1:n}$ to X_{n+1} and is not a "real world variable". Usually we focus on predicting some 'feature of X_{n+1}', say $h(X_{n+1})$ for some function $h : \mathcal{X} \to \mathbb{R}$. The Bayes predictor $\hat{h} = \delta(X_{1:n})$ is the minimizer of the expected loss,

$$\begin{aligned} r(\hat{h}) = \mathbb{E}\rho(\hat{h}(X_{1:n}), h(X_{n+1})) &= \iint_{\mathcal{X}^n\times\mathcal{X}} \rho(\hat{h}(x_{1:n}), x_{n+1}) P(\mathrm{d}x_{1:n}) P(\mathrm{d}x_{n+1}) \\ &= \iiint_{\Theta\times\mathcal{X}^n\times\mathcal{X}} \rho(\hat{h}(x_{1:n}), x_{n+1}) P_\theta(\mathrm{d}x_{1:n}) P_\theta(\mathrm{d}x_{n+1}) \Pi(\mathrm{d}\theta). \end{aligned}$$

The loss function ρ quantifies "how well $\hat{h}(x_{1:n})$ predicts $h(x_{n+1})$" and plays the same role as ℓ in estimation problems. Actually, typical choices of ρ are exactly analogous. For example, the square loss is

$$\rho(a, h(x_{n+1})) = (h(x_{n+1}) - a)^2.$$

Similarly, we can use weighted square loss, precautionary loss, etc. in predicton. However, even if we use "the same loss", the 'best predictor' is in general different from the 'best estimator'.

3.4.6 REMARK. Let $g(\theta) = \mathbb{E}(h(X_i)|\theta)$. Compare estimation of $g(\theta)$ with prediction of $h(X_{n+1})$. If we use the square loss then the Bayes estimator g^* is equal to the Bayes predictor h^*. It follows from the iterated expectation formula: $\hat{h} = \mathbb{E}(h(X_{n+1})|X_{1:n}) = \mathbb{E}(\mathbb{E}(h(X_{n+1})|X_{1:n}, \theta)|X_{1:n}) = \mathbb{E}(\mathbb{E}(h(X_{n+1})|\theta)|X_{1:n}) = \mathbb{E}(g(\theta)|X_{1:n}) = \hat{g}$, in view of the conditional independence $X_{1:n} \perp\!\!\!\perp X_{n+1}|\theta$.

On the other hand, the MSE of prediction is always bigger than that of estimation. It is easy to see that

$$\mathbb{E}(\hat{g} - g(\theta))^2 = \mathbb{E}\mathrm{Var}(g(\theta)|X_{1:n}),$$
$$\mathbb{E}(\hat{h} - h(X_{n+1}))^2 = \mathbb{E}\mathrm{Var}(g(\theta)|X_{1:n}) + \mathbb{E}\mathrm{Var}(h(X_{n+1})|\theta).$$

3.4.7 EXAMPLE. Consider the Normal/Normal model introduced in Example 3.2.3: let $(X_1, \dots, X_n, X_{n+1}|\mu) \sim_{\text{i.i.d.}} \mathrm{N}(\mu, \sigma^2)$ and $\mu \sim \mathrm{N}(m, v^2)$. If we use the square loss then the best estimator of μ and simultaneously the best predictor of X_{n+1} is

$$\hat{X}_{n+1} = \hat{\mu} = z\bar{X} + (1-z)m, \quad \text{where} \quad z = \frac{nv^2}{nv^2 + \sigma^2}.$$

The estimation and prediction MSE is, respectively,

$$\mathbb{E}(\hat{\mu} - \mu)^2 = \frac{\sigma^2 v^2}{nv^2 + \sigma^2},$$
$$\mathbb{E}(\hat{\mu} - X_{n+1})^2 = \frac{\sigma^2 v^2}{nv^2 + \sigma^2} + \sigma^2.$$

△

3.4.8 EXAMPLE (Continuation). Consider the quantile loss (Example 1.1.2) in the same Normal/Normal model. The Bayes estimator of μ and the predictor of X_{n+1} minimize, respectively,

$$\mathbb{E}q_p(\hat{\mu} - \mu)$$
$$\mathbb{E}q_p(\hat{X}_{n+1} - X_{n+1})^2,$$

where $q_p(t) = t(1-p) + t\mathbb{1}(t < 0)$, see Remark 1.1.3. We have

$$\hat{\mu} = z\bar{X} + (1-z)m + \Phi^{-1}(p)\sqrt{\frac{\sigma^2 v^2}{nv^2 + \sigma^2}},$$
$$\hat{X}_{n+1} = z\bar{X} + (1-z)m + \Phi^{-1}(p)\sqrt{\frac{\sigma^2 v^2}{nv^2 + \sigma^2} + \sigma^2},$$

where Φ^{-1} is the quantile function of the standard normal $\mathrm{N}(0, 1)$. △

Similar example with the LINEX loss is in Problem 7.

Problems

1. (Exponential/Gamma) Assume that $X_1, \ldots, X_n, X_{n+1}|\theta \ \sim_{\text{i.i.d.}} \text{Ex}(\theta)$ and the prior is $\theta \sim \text{Gamma}(\alpha, \lambda)$.

   Compute the posterior distribution $\theta|x_{1:n}$.

   Compute the marginal distribution of X_1.

   What is the predictive distribution $X_{n+1}|x_{1:n}$?

2. (Power/Gamma) Random variables $X_1, \ldots, X_n$ are, given θ, conditionally independent with density

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{for } 0 < x < 1; \\ 0 & \text{otherwise.} \end{cases}$$

   Parameter θ has prior distribution Gamma(α, λ).

   Compute the posterior $\pi(\theta|x_1, \ldots, x_n)$.

   Give a one-dimensional sufficient statistic.

   Compute Bayesian estimate of θ for quadratic loss function.

3. (Pareto/Gamma) Random variables $X_1, \ldots, X_n$ are, given θ, conditionally independent with density

$$f(x|\theta) = \begin{cases} \theta(1+x)^{-\theta-1} & \text{for } x \geq 0; \\ 0 & \text{for } x < 0 \end{cases}$$

   Parameter θ has prior distribution Gamma(α, λ).

   Compute the posterior $\pi(\theta|x_1, \ldots, x_n)$.

   Give a one-dimensional sufficient statistic.

   Compute Bayesian estimate of θ for quadratic loss function.

4. (Normal/IG) Compute the marginal distribution in Example 3.2.4 (the result is rescaled noncentral Student's t).

5. (Normal/Normal/IG) Compute the posterior in Example 3.2.5.

6. (Dirichlet and Gamma distributions) Let $\xi_1, \ldots, \xi_d$ be independent r.v.s with $\xi_i \sim \text{Gamma}(\alpha_i, \lambda)$, $S = \xi_1 + \cdots + \xi_d$ and

$$(\vartheta_1, \ldots, \vartheta_d) = \left(\frac{\xi_1}{S}, \ldots, \frac{\xi_d}{S}\right).$$

   Compute the joint distribution of random vector ϑ.

Show that $\vartheta \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_d)$.

Show that ϑ is independent of S.

Hint: Compute the joint density of $(\vartheta_1, \ldots, \vartheta_{d-1}, S)$.

7. Consider the Normal/Normal model (Example 3.2.3). Compute the Bayes estimator of μ and the Bayes predictor of X_{n+1} using the LINEX function: minimize $\mathbb{E}l(\hat{\mu} - \mu)$ and $\mathbb{E}l(\hat{X}_{n+1} - X_{n+1})$, where $l(t) = \mathrm{e}^{\kappa t} - \kappa t - 1$.

8. Let $S = \sum_{i=1}^{N_1} X_i$, where $X_1, \ldots, X_i, \ldots$ are i.i.d. with a known distribution and $N_1 \sim \mathrm{Poiss}(t_1\theta)$ is independent of all X_i. We observe $N_0 = n_0$, where $N_1 \perp\!\!\!\perp N_0|\theta$ and $N_0 \sim \mathrm{Poiss}(t_0\theta)$.

   Compute the Bayes predictor of S with respect to the square loss.

   Compute the Bayes predictor of S with respect to the LINEX loss.

   Hint: The second predictor is expressed in terms of the m.g.f. $M(r) = \mathbb{E}\mathrm{e}^{rX_i}$.

3.5 Classification and prediction

Assume that an 'object' belongs to one of the 'classes' labelled $1, \ldots, k$, but we do not know to which it belongs. We observe a vector $x = (x_1, \ldots, x_d) \in \mathcal{X} \subseteq \mathbb{R}^d$, whose components describe 'features' of the object. On the basis of x we are to 'classify' the object, that is to guess its class assignment. For example, the 'object' may be a patient, the classes correspond to (mutually exclusive) diagnoses and compnents of x may be symptoms, results of diagnostic medical tests etc.

If we assume that the (unobserved) class label c and the (observed) feature vector x are values of random variables C and X, then the classification problem fits in the Bayesian setting. The set of class labels $\mathcal{C} = \{1, \ldots, k\}$ can be treated as a (particulary simple, because finite) space of parameters. The role of likelihood is played by the collection of 'intra-class' densities $\{p(x|c), c = 1, \ldots, k\}$ which describe the probability distributions of the feature vector X, given $C = c$. The prior probabilities $\pi(c)$ describe the relative frequencies of class occurrences. Note that the prior distribution has an objecive interpretation. In our medical diagnosis example, $\pi(c)$ can be identified with the percentage of patients suffering from cth disease (in a hypothetical population of potential patients).

The natural space of decisions in the classification problem is either $\mathcal{A} = \mathcal{C}$ or $\mathcal{A} = \mathcal{C} \cup \{0\}$. After observing x, we guess that the class assignment is $\hat{c}(x) \in \mathcal{C}$ or, perhaps, we say "I don't know". The last action – deferred decision – is encoded as $\hat{c}(x) = 0$. A decision rule $\hat{c} : \mathcal{X} \to \mathcal{A}$ is called a *classifier*. It is convenient to describe $\hat{c}$ as a partition of the feature space $\mathcal{X}$ into a sum of disjoint decision regions $\mathcal{D}_a = \{x : \hat{c}(x) = a\}$. The loss function can, in principle, be any matrix $(\ell(a, c), a \in \mathcal{A}, c \in \mathcal{C})$ such that

- $\ell(a, c) \geq 0$ for all a and c,
- $\ell(c, c) = 0$ for all c (we lose nothing if we guess correctly),
- for all $a \neq 0$ we have $\ell(a, c) > \ell(0, c)$ for at least one c (we loose less if we defer decision than if we make a wrong decision).

From Theorem 2.2.5 it follows that the the best decision rule, *the Bayes classifier* $c^* : \mathcal{X} \to \mathcal{A}$, is given by

$$c^*(x) = \arg\min_a r_x(a) = \arg\min_a \sum_{c=1}^{k} \ell(a, c)\pi(c|x),$$

where the posterior probabilities are computed via the Bayes formula

$$\pi(c|x) = \frac{p(x|c)\pi(c)}{p(x)}.$$

Equivalently,

$$c^*(x) = \arg\min_a \sum_{c=1}^{k} \ell(a, c)p(x|c)\pi(c).$$

If the minimum is attained for more than one a, the way of "resolving a tie" is arbitrary.

In the sequel we focus on the special loss functions. If we permit deferred decision '0' then

$$\ell(a, c) = \begin{cases} 0 & \text{if } a = c; \\ 1 & \text{if } a \neq c \text{ and } a \neq 0; \\ \lambda & \text{if } a = 0. \end{cases}$$

The Bayes risk is

$$r(\hat{c}) = \mathbb{P}(\hat{c}(X) \neq C, \hat{c}(X) \neq 0) + \lambda\mathbb{P}(\hat{c}(X) = 0).$$

The Bayes classifier $c^* : \mathcal{X} \to \mathcal{C} \cup \{0\}$ is given by

$$c^*(x) = \begin{cases} \arg\max_c \pi(c|x) & \text{if } \max_c \pi(c|x) \geq 1 - \lambda; \\ 0 & \text{if } \max_c \pi(c|x) < 1 - \lambda. \end{cases}$$

In the case $\mathcal{A} = \mathcal{C}$ (deferred decision is not permitted), if we let $\ell(a, c) = \mathbb{1}(a \neq c)$ then the Bayes risk is the probability of misclassification,

$$r(\hat{c}) = \mathbb{P}(\hat{c}(X) \neq C).$$

The Bayes classifier $c^* : \mathcal{X} \to \mathcal{C}$ is given by

$$c^*(x) = \arg\max_c \pi(c|x) = \arg\max_c p(x|c)\pi(c) = \arg\max_c \left[\log p(x|c) + \log \pi(c)\right],$$

and usually we use the logarithmic version.

3.5.1 EXAMPLE. [Normal distributions with unequal covariance matrices] Assume that the class-conditional distributions of the feature vector are multivariate normal, $(X|C = c) \sim \mathrm{N}(\mu_c, \Sigma_c)$. Assume that the matices Σ_c are nonsigular. Then

$$p(x|c) = (2\pi)^{-\frac{d}{2}} (\det \Sigma_c)^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}(x-\mu_c)^\top \Sigma_c^{-1}(x-\mu_c)\Big).$$

Consider classification without deferred decision, $\mathcal{A} = \mathcal{C}$. Since we have

$$\log p(x|c) = -\frac{1}{2}(x-\mu_c)^\top \Sigma_c^{-1}(x-\mu_c) - \frac{1}{2}|\det \Sigma_c| + \text{const},$$

the Bayes classifier chooses the maximum of k *quadratic* functions $d_c(x) = \log p(x|c) + \log \pi(c)$. They are called *quadratic discriminant functions*. In fact we need only $k-1$ of them, for example $d_c(x) - d_1(x)$ for $c = 2, \ldots, k$. The borders of Bayes decision regions are quadratic manifolds as in Figure 3.2. △

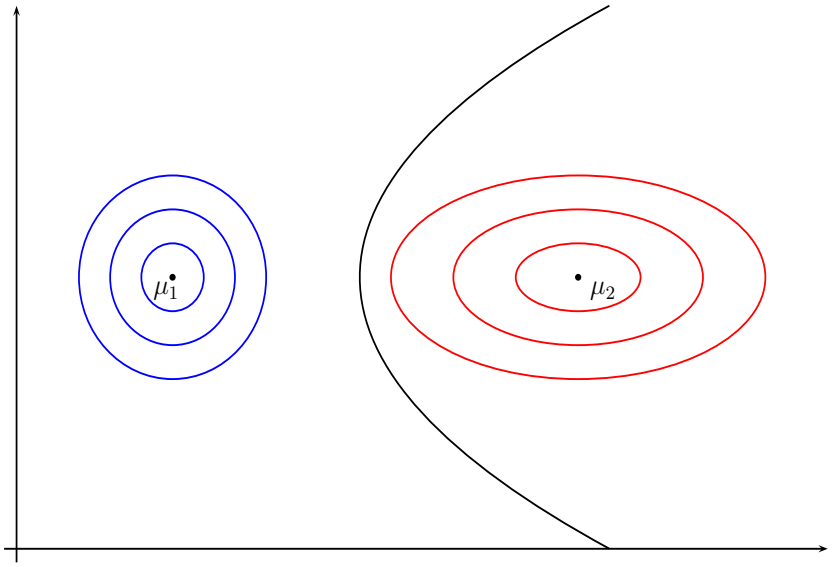


Figure 3.2: Decision regions for QDA (Example 3.5.1).

3.5.2 EXAMPLE. [Normal distributions with equal covariance matrices] Assume additionally that the covariance matrices within classes are equal, $X|C = c \sim \mathrm{N}(\mu_c, \Sigma)$. Now

$$\log p(x|c) - \log p(x|1) = \left(x - \frac{\mu_c + \mu_1}{2}\right)^\top \Sigma^{-1}(\mu_c - \mu_1),$$

and we have $k-1$ *linear discriminant functions* $d_c(x) = \log p(x|c) - \log p(x|c) + \log(\pi(c)/\pi(1))$. The boundaries of Bayes decision regions are hyperplanes as in Figure 3.3. △

In reality this simple model is only the starting point for more complicated considerations. Usually the parameters μ_c and Σ_c (or μ_c and Σ) are unknown and have to be estimated.

Problems

1. (Logistic Regression) Consider the normal model with equal covariance matrices, as in Example 3.5.2, with two classes. Show that the posterior is a logistic function of a linear function of x:
$$p(2|x) = \frac{\exp(\alpha + \beta^\top x)}{1 + \exp(\alpha + \beta^\top x)}$$
for some $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$.

2. (Mahalanobis Distance) In the LDA model of Example 3.5.2, define the Mahalanobis distance between $\xi_1, \xi_2 \in \mathbb{R}^d$ as
$$|\xi_1 - \xi_2|_{\Sigma^{-1}} = \sqrt{(\xi_1 - \xi_2)^\top \Sigma^{-1} (\xi_1 - \xi_2)}.$$
Show that for equal prior probabilities, the Bayes classifier assigns x to the class j for which $|x - \mu_j|_{\Sigma^{-1}}$ is minimum.

3. (Misclassification Probability) In the LDA model with two classes, write the formula for the misclassification probability in terms of the Mahalanobis distance $|\mu_2 - \mu_1|_{\Sigma^{-1}}$ and using the standard normal c.d.f. Φ.

4. (Naive Bayes Classifier for binary features) Assume that $\mathcal{X} = \{0, 1\}^d$ (we observe d binary features) and
$$p(x|j) = \prod_{i=1}^{d} \theta_{ji}^{x_i} (1 - \theta_{ji})^{1 - x_i}.$$
Show that the Bayes classifier is linear (the boundaries of Bayes decision regions are hyperplanes). The name "naive" refers to the simplistic assumption that within each class, the features are independent. Hoverver, this "naive" assumptions dramatically reduce the number of parameters, which are in practice unknown and have to be estimated from a training sample (there are kd parameters θ_{ji}).

3.6 Testing statistical hypotheses

Two simple hypotheses

The problem of testing a null hypothesis against an alternative resembles, at the first glance, classification with two classes. In fact, the abstract formulation is almost identical, but the

context and interpretation is quite different. To begin with, consider two simple hypotheses

$$\mathrm{H}_0 : X \sim p_0,$$
$$\mathrm{H}_1 : X \sim p_1.$$

We can consider the parameter space $\{0,1\}$ and the decision space $\{0,1\}$ (where 1 is interpreted as rejecion of H_0 in favour of H_1). A decision rule $\delta : \{0,1\} \to \{0,1\}$ is a (nonrandomized) test. A prior distribution is given by $\pi(0) = \mathbb{P}(\mathrm{H}_0)$ and $\pi(1) = \mathbb{P}(\mathrm{H}_1)$. The Bayes risk is

$$r(\delta) = \alpha(\delta)\ell_0\pi(0) + \beta(\delta)\ell_1\pi(1),$$

where $\alpha(\delta)$ and $\beta(\delta)$ are the probabilities of two types of error:

$$\alpha(\delta) = \mathbb{P}(\delta(X) = 1|\mathrm{H}_0) = \int \mathbb{1}(\delta(x) = 1)p_0(x)\mathrm{d}x \quad \text{(error of the I type)},$$
$$\beta(\delta) = \mathbb{P}(\delta(X) = 0|\mathrm{H}_1) = \int \mathbb{1}(\delta(x) = 0)p_1(x)\mathrm{d}x \quad \text{(error of the II type)},$$

and $\ell_0 = \ell(1,0) > 0$, $\ell_1 = \ell(0,1) > 0$ are weights corresponding to the two types of error.

The test δ^* is Bayes iff it satisfies

$$\begin{cases} \delta^*(x) = 1 & \text{if} \quad \dfrac{p_1(x)}{p_0(x)} > \dfrac{\ell_0\pi(0)}{\ell_1\pi(1)}; \\ \\ \delta^*(x) = 0 & \text{if} \quad \dfrac{p_1(x)}{p_0(x)} < \dfrac{\ell_0\pi(0)}{\ell_1\pi(1)}. \end{cases}$$

If $p_1(x)/p_0(x) = (\ell_0\pi(0))/(\ell_1\pi(1))$ then the decision $\delta^*(x)$ can be chosen arbitrarily. A Bayes test always exists, but is not necessarily unique.

It is interesting to compare the Bayes test with the MPT (most powerful test) at the given level of significance α_0, advocated by the frequentists. The Neyman-Pearson lemma says that the MPT is *basically* a function of the likelihood ratio, $p_1(x)/p_0(x)$. The Bayes test δ^* is always MPT at the significance level $\alpha_0 = \alpha(\delta^*)$. On the other hand, a nonrandomized MPT is not necessarily Bayes and may be completely 'stupid'. An example is given in Problem 1. A remedy is to randomize decisions. A a randomized test is $\delta_{\mathrm{rand}} : \{0,1\} \to [0,1]$. If $\delta(x) = p$ then we perform a random experiment (a Bernoulli trial with probability of success p). 'Success' is interpreted as decision '1', that is rejection of the null hypothesis. The the probabilities of error for a randomized test are:

$$\alpha(\delta_{\mathrm{rand}}) = \mathbb{E}(\delta_{\mathrm{rand}}(X)|\mathrm{H}_0) = \int \delta_{\mathrm{rand}}(x)p_0(x)\mathrm{d}x \quad \text{(error of the I type)},$$
$$\beta(\delta_{\mathrm{rand}}) = 1 - \mathbb{E}(\delta_{\mathrm{rand}}(X)|\mathrm{H}_1) = 1 - \int \delta_{\mathrm{rand}}(x)p_1(x)\mathrm{d}x \quad \text{(error of the II type)}.$$

Define the 'risk set' $\mathcal{R} \subseteq [0,1]^2$ by

$$\mathcal{R} = \{(\alpha(\delta_{\text{rand}}), \beta(\delta_{\text{rand}}) : \delta_{\text{rand}} \text{ is a randomized test. } \}$$

$\mathcal{R}$ has the following nice properties.

- $\mathcal{R}$ is a convex and closed set,
- $(0,1) \in \mathcal{R}$ and $(1,0) \in \mathcal{R}$,
- if $(\alpha, \beta) \in \mathcal{R}$ then $(1-\alpha, 1-\beta) \in \mathcal{R}$.

Using these facts, it is easy to show that every randomized MPT δ_{rand} at some significance level $\alpha_0 > 0$ is Bayes for some $\ell_0, \ell_1 > 0$, provided that $\alpha(\delta_{\text{rand}}), \beta(\delta_{\text{rand}}) > 0$. (As we are Bayesians, we keep $\pi(0), \pi(1) > 0$ fixed.) See Problems 2, 3. Note the analogy with Propositions 2.2.10 and 2.2.11.

Problems

1. Consider two probability distributions on space $\mathcal{X} = \{1,2,3\}$:

x	1	2	3
$p_0(x)$	0.2	0.5	0.3
$p_1(x)$	0.1	0.3	0.6

   Consider the problem of testing $\text{H}_0 : X \sim p_0$ vs $\text{H}_1 : X \sim p_1$.

   Sketch the set $\mathcal{R}_{\text{non-rand}} = \{\alpha(\delta), \beta(\delta) : \delta \text{ is a nonrandomized test}\}$. What is the nonrandomized MPT at the level of significance $\alpha_0 = 0.25$?

   Sketch the set $\mathcal{R} = \{(\alpha(\delta_{\text{rand}}), \beta(\delta_{\text{rand}}) : \delta_{\text{rand}} \text{ is a randomized test}\}$. What is the randomized MPT at the level of significance $\alpha_0 = 0.25$?

2. Show that the risk set $\mathcal{R}$ is a convex and closed set.

   Hint: It is *not* trivial to verify rigorously the second property.

3. Show that every randomized MPT δ_{rand} at some significance level $\alpha_0 > 0$ is Bayes for some $\ell_0, \ell_1 > 0$, provided that $\alpha(\delta_{\text{rand}}), \beta(\delta_{\text{rand}}) > 0$. Find a counter-example to this statement for non-randomized tests.

4. Consider the problem of testing $\text{H}_0 : X \sim \text{Ex}(1)$ vs $\text{H}_1 : X \sim \text{Ex}(1) + 1$, where $\text{Ex}(1)$ is the exponential distribution with parameter 1 and '$\text{Ex}(1) + 1$' is the exponential shifted by 1 to the right $p_1(x) = \text{e}^{-(x-1)} \mathbb{1}(x > 1)$. Make a graph of the risk set $\mathcal{R}$.

5. Consider the problem of testing $\text{H}_0 : X \sim \text{N}(0,1)$ vs $\text{H}_1 : X \sim \text{N}(\mu_1, 1)$, where μ_1 is fixed. Make a graph of the risk set $\mathcal{R}$. For the MPT at the significance level $\alpha > 0$, find loss coefficients $\ell_0, \ell_1 > 0$ which make this test Bayes (Problem 3).

Composite hypotheses and Bayes factors

Let us reformulate the problem of testing statistical hypotheses as the problem of *model choice*. To begin with, consider two Bayesian models, $M = 0$ and $M = 1$, with two likelihood functions and two priors (possibly on two different parameter spaces). If we are consitent Bayesians, we also ought to define prior probabilities of the two models. Summing up, we have the following elements:

$$\begin{aligned} &p(x|\theta_0, M = 0), \qquad \pi_0(\theta_0), \qquad \pi(0) = \mathbb{P}(M = 0), \\ &p(x|\theta_1, M = 1), \qquad \pi_1(\theta_1), \qquad \pi(1) = \mathbb{P}(M = 0). \end{aligned}$$

(In most interesting examples the parameter space in model 0 is a subset of that in model 1; then symbol θ is used instead of θ_0 and θ_1.) Upon observing $X = x$, we are to decide which is true:

$$\mathrm{H}_0 : M = 0 \quad \text{or} \quad \mathrm{H}_1 : M = 1.$$

Bayes formula allows us to compute $\mathbb{P}(M = i|x)$ for $i = 0, 1$:

$$\begin{aligned} \mathbb{P}(M = 1|x) &= \frac{\pi(1)p(x|M = 1)}{\pi(1)p(x|M = 1) + \pi(0)p(x|M = 0)} \\ &= \frac{\pi(1)p(x|M = 1)/p(x|M = 0)}{\pi(1)p(x|M = 1)/p(x|M = 0) + \pi(0)}. \end{aligned}$$

The posterior probabilities of the hypotheses (models) depend on $\pi(i)$ and the quotient

$$B_{10} = \frac{p(x|M = 1)}{p(x|M = 0)},$$

named *Bayes Factor* (BF). It is interpreted as the evidence in favour of H_1 against H_0. The standard approach is to compute the BF without specifying the $\pi(i) = \mathbb{P}(\mathrm{H}_i)$. (The default decision threshold is $B_{10} = 1$, because the Bayes decision rule is "reject of H_0 in favour of H_1 when $B_{10} > 1$", provided that $\pi(0) = \pi(1) = 0.5$ and the loss function is symmetric $\ell(0, 1) = \ell(1, 0)$. However, usually it is impossible or undesirable to choose $\pi(i)$ and ℓ. The threshold $B_{10} = 1$ plays in Bayesian statistics analogous role as the "standard level of significance 0.95" in frequentist statistics. The default frequentist rule is to "reject of H_0 in favour of H_1 when the p-value is < 0.95".)

Typically, the competing models are of different "dimension" and we are to decide whether to choose a "smaller" or "bigger" model (i.e. a model with fewer or more parameters. The simplest case is testing a simple H_0 versus a composite H_1, i.e. a fully specified model versus a model with unknown parameters, i.e. dim $= 0$ vs dim > 0.

The approach based on Bayes Factors generalizes to the problem of choosing among more than two competing models. If we have k models ($M = 0, 1, \ldots, k - 1$) then

$$\mathbb{P}(M = i|x) = \frac{\pi(i)B_{i0}}{\pi(0) + \sum_{l=1}^{k-1} \pi(l)B_{l0}}.$$

The Bayes factor is the ratio of *marginal likelihoods* in the competing models:

$$p(x|M=i) = \int p(x|\theta_i, M=i)\pi_i(\theta_i)\mathrm{d}\theta_i.$$

Apart from simple conjugate models, the *marginal likelihoods* and consequently the BFs are difficult to compute. The following example describes a rather rare situation where there is an explicit formula for the BF.

3.6.1 EXAMPLE. Let $X_1, \ldots, X_n$ be a sample from $\mathrm{N}(\theta, \sigma^2)$. Assume that σ^2 is known and consider the following two competing hypotheses about θ. The null hypothesis fully specifies the parameter value, $\mathrm{H}_0 : \theta = \mu$, where μ is given. The alternative is $\mathrm{H}_1 : \theta \sim \mathrm{N}(\mu, v^2)$. (Note that the same μ appears in the null and in the alternative.) To compute the bayes Factor, use the marginal densities of the sufficient statistic $\bar{x} = \sum x_i/n$. Since

$$p(\bar{x}|\mathrm{H}_0) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-\frac{1}{2\sigma^2/n}(\bar{x}-\mu)^2\right),$$
$$p(\bar{x}|\mathrm{H}_1) = \frac{1}{\sqrt{2\pi\left(\sigma^2/n + v^2\right)}} \exp\left(-\frac{1}{2\left(\sigma^2/n + v^2\right)}(\bar{x}-\mu)^2\right),$$

we obtain

$$B_{10} = \frac{p(\bar{x}|\mathrm{H}_1)}{p(\bar{x}|\mathrm{H}_0)} = \sqrt{\frac{\sigma^2/n}{\sigma^2/n + v^2}} \exp\left(\frac{v^2}{2\left(\sigma^2/n + v^2\right)\sigma^2/n}(\bar{x}-\mu)^2\right).$$

A Bayes test rejects H_0 in favour of H_1 if the statistic $z = \sqrt{n}|\bar{x}-\mu|/\sigma$ exceeds some threshold value c, similarly as a frequentists' test on a given level of significance. The difference is in the choice of c for both tests. This difference leads to a phenomenon known as "Jeffreys-Lindley's Paradox", see Problem 1. △

In the above example we used the following proposition.

3.6.2 Proposition. *Consider the problem of testing a simple null hypothesis* $\mathrm{H}_0 : \theta = \theta_0$ *against a "diffuse" alternative* $\mathrm{H}_1 : \theta \sim \pi(\cdot)$, *where* $\theta_0 \in \Theta$ *is some fixed value and* π *is some prior distribution over* Θ. *Assume that* $T = T(X)$ *is a sufficient statistic under* H_1 *and* $p(t|\theta)$ *is the density of this statistic at* $t = T(x)$. *Then*

$$B_{10} = \frac{\int_\Theta p(t|\theta)\pi(\theta)\mathrm{d}\theta}{p(t|\theta_0)}.$$

Proof. If $T(x) = t$ then $p(x|\theta) = p(x|t,\theta)p(t|\theta) = p(x|t)p(t|\theta)$, because $X \perp\!\!\!\perp \theta | T$. Consequently,

$$B_{10} = \frac{\int_\Theta p(x|\theta)\pi(\theta)\mathrm{d}\theta}{p(x|\theta_0)} = \frac{p(x|t)\int p(t|\theta)\mathrm{d}\theta}{p(x|t)p(t|\theta_0)}.$$

□

Problems

1. (Jeffreys-Lindley's paradox) Reconsider Example 3.6.1. Let $z_n^2 = n(\bar{x} - \mu)^2/\sigma^2$ be the standard χ^2 statistic for testing $\mathrm{H}_0 : \theta = \mu$ against the alternative $\theta \neq \mu$ in the frequentist setting. Note that the Bayes factor computed in Example 3.6.1 is also an increasing function of z_n^2. Compare the frequentist test with the Bayesian test based on B_{01}. Asssume that $z_n^2 = 2$ is constant as $n \to \infty$. It is obvious that the frequentist p-value remains constant, and leads to rejection of H_0 (if, for example, the significance level is $\alpha = 0.05$). On the other hand, show that $B_{10} \to 0$ with $n \to \infty$, so a Bayesian has no reason to reject H_0!

2. Consider a family of binomial distributions $\mathrm{Bin}(n, \theta)$ and the problem of testing $\mathrm{H}_0 : \theta = \gamma$ against $\mathrm{H}_1 : \theta \sim \mathrm{Beta}(s\gamma, s(1-\gamma)$. Compute the Bayes factor B_{10} (do not expect a very simple expression).

   Remark: Note that under H_1 we have $\mathbb{E}\theta = \gamma$, so the alternative is a sort of "diffused null", similarly as in Example 3.6.1.

3.7 Credible Regions

The Bayesian counterpart of "interval estimation" is particularly simple and straightforward. Instead of "confidence intervals", in the Bayesian world we have credible intervals or more generally sets (regions). By definition a set $C_x \subseteq \Theta$ is a *credible region* at the level $1 - \alpha$ if

$$\mathbb{P}(\theta \in C_x | x) = \int_{C_x} \Pi_x(\mathrm{d}\theta) \geq 1 - \alpha.$$

As usual in the Bayesian theory, it is easy to define some notion of optimality.

3.7.1 Lemma. *Let $\pi(\cdot)$ be a probability density on Θ. For $h > 0$ let $C^h = \{\theta : \pi(\theta) \geq h\}$. If for some set $C \subseteq \Theta$ we have*

$$\int_{C^h} \pi(\theta)\mathrm{d}\theta \leq \int_C \pi(\theta)\mathrm{d}\theta$$

then

$$\int_{C^h} \mathrm{d}\theta \leq \int_C \mathrm{d}\theta.$$

Proof. By assumption

$$\int_{C^h \setminus C} \pi(\theta)\mathrm{d}\theta \leq \int_{C \setminus C^h} \pi(\theta)\mathrm{d}\theta.$$

On the set $C^h \setminus C$ we have $\pi \geq h$, while on $C \setminus C^h$ the opposite, $\pi < h$. Therefore

$$\int_{C^h} h \mathrm{d}\theta \leq \int_C h \mathrm{d}\theta.$$

It is enough to divide both sides by h and add $\int_{C^h \cap C} \mathrm{d}\theta$. □

Of course, this lemma applies to the posterior density as well as to the prior. The set $C_x^h = \{\theta : \pi_x(\theta) \geq h\}$ is called Highest Posterior Density (HPD) Region. The integral $\int_C \mathrm{d}\theta$ is the 'volume' of set C. We obtain the following fact.

3.7.2 Proposition (Highest Posterior Density Regions)**.** *If for some $h > 0$ the set C_x^h is a HDP region such that $\mathbb{P}(\theta \in C_x^h | x) = 1 - \alpha$ and for some set C_x we have $\mathbb{P}(\theta \in C_x | x) \geq 1 - \alpha$ then the volume of C_x^h is less than that of C_x.*

Put differently, the HDP is the "smallest" credible region (with respect to the 'volume' i.e. the base measure $\mathrm{d}\theta$).

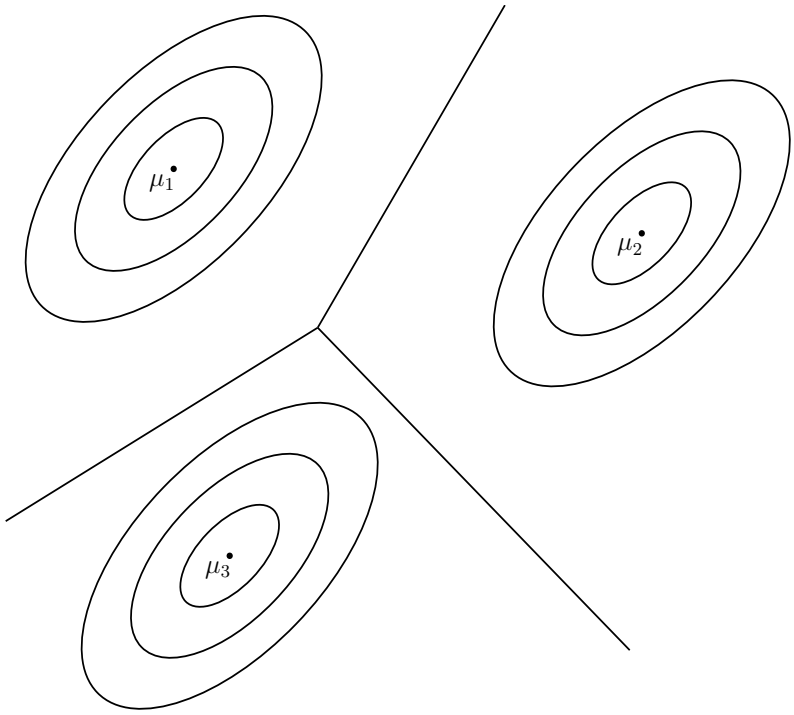


Figure 3.3: Decision regions for LDA (Example 3.5.2).

Chapter 4

Monte Carlo methods in Bayesian computations

4.1 Hierarchical models and Gibbs Sampler

Model of variance components

The following model is useful in various fields, including survey statistics, actuarial mathematics and agriculture. Suppose that the data are values of a scalar, continuous variable for 'units' which belong to different 'groups'. For example, in survey statistics the variable can be the income, 'units' can be firms and 'groups' can correspond to branches of business. The object of statistical inference in this example is the vector of branch-specific mean amounts of income. The main idea behind a statistical model below is that the variability within groups is smaller than variability between groups.

- $y_{ij} = Y_{ij} \sim_{\text{i.i.d.}} \mathrm{N}(\theta_i, \sigma^2)$ – observed value for jth 'unit' in ith group, $(j = 1, \ldots, n_i)$, $(i = 1, \ldots, k)$,
- $\theta_i \sim_{\text{i.i.d.}} \mathrm{N}(\mu, \upsilon^2)$ – the value of interest (mean in ith group),
- μ – global mean.

This is a basic model of variance components. It follows from Example 3.2.3 that the Bayes estimators for the square loss are

$$\hat{\theta}_i = \mathbb{E}(\theta_i|y) = z_i \bar{y}_i + (1 - z_i)\mu, \quad \text{where} \quad z_i = \frac{n_i \upsilon^2}{n_i \upsilon^2 + \sigma^2}, \quad \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}.$$

Moreover, Example 3.4.7 shows that $\hat{\theta}_i$ is also the best predictor of the future observation belonging to ith group.

The problem is that $\hat{\theta}_i$ depends on three parameters: μ, σ^2 and υ^2, which in practice are unknown and have to be estimated from data. A consistently Bayesian solution is to introduce prior distributions on these parameters and thus construct a higher layer of hierarchy. If we use conjugate priors then the extended hierarchical model is the following.

- $Y_{ij} \sim \mathrm{N}(\theta_i, \sigma^2)$,
- $\theta_i \sim \mathrm{N}(\mu, \upsilon^2)$,
- $\mu \sim \mathrm{N}(m, \tau^2)$,
- $\sigma^{-2} \sim \mathrm{Gamma}(p, \kappa)$,
- $\upsilon^{-2} \sim \mathrm{Gamma}(q, \lambda)$.

The joint density of all model variables is thus

$$p(y, \theta, \mu, \sigma^{-2}, \upsilon^{-2}) = p(y|\theta, \sigma^{-2})\pi(\theta|\mu, \upsilon^{-2})\pi(\mu)\pi(\sigma^{-2})\pi(\upsilon^{-2}).$$

(To avoid introducing new notation, let us adopt the convention that symbols υ^{-2} and σ^{-2} denote 'new' parameters, and are not treated as functions of υ and σ.) The joint posterior is given by

$$\pi(\theta, \mu, \sigma^{-2}, \upsilon^{-2}|y) = \frac{p(y, \theta, \mu, \sigma^{-2}, \upsilon^{-2})}{p(y)}.$$

This probability distribution is intractable, mainly because of the norming constant $p(y)$, which is a an integral over a subset of $\mathbb{R}^{k+3}$. However all the full conditionals are simple. If we explicitly write the posterior density, single out one of the variables and consider only expressions containing this variable, we easily obtain the full conditional (due to conjugacy of priors). For example, below we focus on υ^{-2}. Only the expressions indicated in blue contain υ^{-2}, the rest can be treated as constant.

$$\begin{aligned}\pi(\theta, \mu, \upsilon^{-2}, \sigma^{-2}|y) \propto (\sigma^{-2})^{n/2} \exp\left\{-\frac{\sigma^{-2}}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2\right\} \\ \cdot (\upsilon^{-2})^{k/2} \exp\left\{-\frac{\upsilon^{-2}}{2}\sum_{i=1}^{k}(\theta_i-\mu)^2\right\} \\ \cdot \exp\left\{-\frac{\tau^{-2}}{2}(\mu-m)^2\right\} \\ \cdot (\sigma^{-2})^{q-1}\exp\{-\kappa\sigma^{-2}\} \\ \cdot (\upsilon^{-2})^{p-1}\exp\{-\lambda\upsilon^{-2}\}.\end{aligned}$$

Consequently,

$$\pi(\upsilon^{-2}|y,\theta,\mu,\sigma^{-2}) \propto (\upsilon^{-2})^{k/2+p-1} \cdot \exp\left\{-\left(\frac{1}{2}\sum_{i=1}^{k}(\theta_i-\mu)^2+\lambda\right)\upsilon^{-2}\right\}.$$

The full conditional distribution of υ^{-2} is Gamma$(k/2+p, \sum_{i=1}^{k}(\theta_i-\mu)^2/2+\lambda)$. Analogously we compute the other full conditionals:

$$\begin{aligned}
\upsilon^{-2}|y,\theta,\mu,\sigma^{-2} &\sim \text{Gamma}\left(\frac{k}{2}+p, \frac{1}{2}\sum_{i=1}^{k}(\theta_i-\mu)^2+\lambda\right),\\
\sigma^{-2}|y,\theta,\mu,\upsilon^{-2} &\sim \text{Gamma}\left(\frac{n}{2}+q, \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2+\kappa\right),\\
\mu|y,\theta,\sigma^{-2},\upsilon^{-2} &\sim \text{N}\left(\frac{k\tau^2}{k\tau^2+\upsilon^2}\bar{\theta}+\frac{\upsilon^2}{k\tau^2+\upsilon^2}m, \frac{\tau^2\upsilon^2}{k\tau^2+\upsilon^2}\right),\\
\theta_i|y,\theta_{-i},\mu,\sigma^{-2},\upsilon^{-2} &\sim \text{N}\left(\frac{n_i\upsilon^2}{n_i\upsilon^2+\sigma^2}\bar{y}_i+\frac{\sigma^2}{n_i\upsilon^2+\sigma^2}\mu, \frac{\upsilon^2\sigma^2}{n_i\upsilon^2+\sigma^2}\right),
\end{aligned}$$

where $n=\sum n_i$, $\bar{\theta}=\sum_i\theta_i/k$ and $\theta_{-i}=(\theta_l)_{l\neq i}$. Note that the components of θ are conditionally independent (θ_i is independent of θ_{-i}). The Gibbs Sampler can update θ as a single "block". The state of the Markov chain is a point $\psi=(\theta,\mu,\sigma^{-2},\upsilon^{-2})\in\mathbb{R}^{k+3}$. The transition rule (one 'big step' of the GS) is

$$\underbrace{\left(\theta,\mu,\sigma^{-2},\upsilon^{-2}\right)}_{\Psi_t} \longmapsto \underbrace{\left(\theta,\mu,\sigma^{-2},\upsilon^{-2}\right)}_{\Psi_{t+1}},$$

and is composed of the following 'small steps':

- Sample $\upsilon^{-2} \sim p(\upsilon^{-2}|y,\theta,\mu,\sigma^{-2}\quad) = \text{Gamma}(\ldots)$,
- Sample $\sigma^{-2} \sim p(\sigma^{-2}|y,\theta,\mu,\quad\upsilon^{-2}) = \text{Gamma}(\ldots)$,
- Sample $\mu \sim p(\mu|y,\theta,\sigma^{-2},\upsilon^{-2}) = \text{N}(\ldots)$,
- Sample $\theta \sim p(\theta|y,\mu,\sigma^{-2},\upsilon^{-2}) = \text{N}(\ldots)$.

Markov chain Ψ_t converges to the posterior:

$$\Psi_t \to \pi(\cdot|y) = \pi(\theta,\mu,\sigma^{-2},\upsilon^{-2}|y).$$

Suppose that the aim of the analysis is estimation/prediction of θ_1, that is computing

$$\mathbb{E}(\theta_1|y) = \int\cdots\int \theta_1\pi(\theta,\mu,\sigma^{-2},\upsilon^{-2}|y)\mathrm{d}\theta_2\cdots\mathrm{d}\theta_k\mathrm{d}\mu\mathrm{d}\sigma^{-2}\mathrm{d}\upsilon^{-2}.$$

The MCMC approximation of these expectation is an average along the trajectory of the chain:

$$\theta_1(\Psi_0), \theta_1(\Psi_1), \ldots, \theta_1(\Psi_t), \ldots$$

where $\theta_1(\psi) = \theta_1$ for $\psi = (\theta_1, \ldots, \theta_k, \mu, \sigma^{-2}, \upsilon^{-2})$.

Model of Gaussian mixtures

As in the model of variance components, we consider observations partitioned into several groups. The difference is that the partition is 'hidden'. We do not know which observations belong to the same group. Data is a sample from a mixture distribution $\sum_{j=1}^k p_j P_{\theta_j}(\cdot)$, where $\sum_{j=1}^k p_j = 1$ and each $P_{\theta_j}(\cdot)$ is a probability distribution. The parameters θ_j are unknown, the component probabilities p_j are unknown. Only the number of components k is assumed to be known. Consider the normal likelihood within the components and the following hierarchy of (conjugate) priors:

- $Y_1, \ldots, Y_n \sim_{\text{i.i.d.}} \sum_{j=1}^k p_j \mathrm{N}(\mu_j, \sigma_j^2)$,
- $(p_1, \ldots, p_k) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$,
- $\mu_1, \ldots, \mu_k \sim_{\text{i.i.d.}} \mathrm{N}(m, v^2)$,
- $\sigma_1^{-2}, \ldots, \sigma_k^{-2} \sim_{\text{i.i.d.}} \mathrm{Gamma}(\gamma, \lambda)$.

The hyperparameters $m, v^2, \beta, \gamma, \lambda$ are fixed and given. The target distribution is the posterior, $\pi(p, \mu, \sigma^{-2}|y)$, where $\sigma^{-2} = (\sigma_1^{-2}, \ldots, \sigma_k^{-2})$.

The GS algorithm is based on the idea of auxillary variables. In our model, we introduce latent variables $c_1, \ldots, c_n$ which indicate from which components come the observations. Explicitly, we assume that

- $\mathbb{P}(c_i = j|p) = p_j$ (*a priori*),
- $Y_i|c_i = j \sim \mathrm{N}(\mu_j, \sigma_j^2)$ independently of the rest of variables.

In general, the role of auxillary variables is to facilitate construction of an MCMC algorithm. In the model of mixtures they are clearly of independent interest.

The joint posterior density in our model is

$$\pi(p, c, \mu, \sigma^{-2}|y) \propto \prod_{i=1}^n p_{c_i} (\sigma_{c_i}^{-2})^{1/2} \exp\left(-\frac{\sigma_{c_i}^{-2}}{2}(y_i - \mu_{c_i})^2\right) \cdot \prod_{j=1}^k p_j^{\alpha_j - 1} \cdot \prod_{j=1}^k \exp\left(-\frac{1}{2v^2}(\mu_j - m)^2\right) \cdot \prod_{j=1}^k (\sigma_j^{-2})^{\gamma-1} \exp\left(-\lambda \sigma_j^{-2}\right).$$

From this formula we easily extract the full conditionals. They are posterior distributions in well-known conjugate models which were considered in Sections 3.2 and 3.5.

- The full conditional of c. Independently for $i = 1, \dots, n$ we have

$$\mathbb{P}(c_i = j | p, \mu, \sigma^{-2}, y) \propto p_j (\sigma_j^{-2})^{1/2} \exp\left(-\frac{\sigma_j^{-2}}{2} (y_i - \mu_j)^2 \right).$$

  This is exactly the problem of classification for n 'objects'. The solution is the QDF described in Section 3.5.

- The full conditional of p. Rearranging expressions with p we obtain

$$\pi(p | c, \mu, \sigma^{-2}, y) \propto \prod_{j=1}^{k} p_j^{\alpha_j + n_j - 1},$$

  where $n_j = \sum_{i=1}^{n} \mathbb{1}(c_i = j)$. The full conditional is thus $\mathrm{Dir}(\alpha_k + n_k, \dots, \alpha_k + n_k)$.

- The full conditional of μ. The part of the posterior depending on μ can be rewritten as

$$\pi(\mu | p, c, \sigma^{-2}, y) \propto \prod_{j=1}^{k} \exp\left(-\frac{\sigma_j^{-2}}{2} \sum_{i: c_i = j} (y_i - \mu_j)^2 \right) \exp\left(-\frac{1}{2v^2} (\mu_j - m)^2 \right).$$

  Independently for $j = 1, \dots, k$, we are to compute the posterior in the Normal/Normal model, as in Example 3.2.3. We obtain the normal distribution,

$$\mu_j \sim \mathrm{N}\left(z_j \bar{y}_j + (1 - z_j) m, \frac{v^2}{n_j \sigma_j^{-2} + 1} \right),$$
$$z_j = \frac{n_j \sigma_j^{-2} v^2}{n_j \sigma_j^{-2} v^2 + 1}, \quad \bar{y}_j = \frac{1}{n_j} \sum_{i: c_i = j} y_i.$$

  Every time we use the current class allocations c_i, which can change from step to step.

- The full conditional of σ^{-2}. We only have to recall the Normal/IG model in Example 3.2.4. The full conditional is

$$\sigma_j^{-2} \sim \mathrm{Gamma}\left(\gamma + \frac{n_j}{2}, \lambda + \frac{1}{2} \sum_{i: c_i = j} (y_i - \mu_j)^2 \right),$$

  independently for every (current) group j.

4.2 Hidden Markov models and sequential Monte Carlo

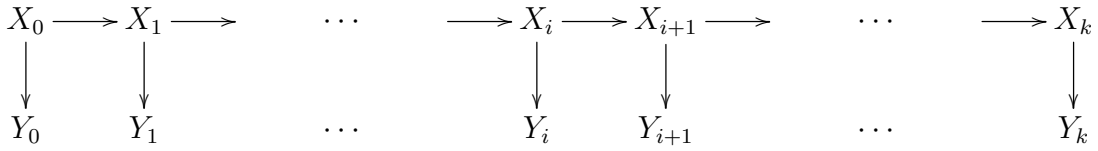
The simplest *Hidden Markov Model* (HMM, also known as "state-space" model) is the following.

- $X_0, X_1, \ldots, X_k$ is a hidden (unobservable) Markov chain with transition matrix $T = (T(x_i, x_{i+1}))$ and initial distribution $s = (s(x_0))$.
- $Y_0, Y_1, \ldots, Y_k$ are observed random variables such that Y_i depends only on X_i and has the likelihood $L(x_i, y_i) = p(y_i|x_i)$.

The model is thus specified by the joint probability distribution:

$$p(x_0, \ldots, x_k, y_0, \ldots, y_k) = s(x_0)T(x_0, x_1) \cdots T(x_i, x_{i+1}) \cdots T(x_{k-1}, x_k) \\ L(x_0, y_0)L(x_1, y_1) \cdots L(x_i, y_i) \cdots L(x_k, y_k).$$

The graphical representation of the model is given below. It indicates the structure of dependencies (or rather conditional independencies) between the variables.



Forward-Backward algorithm

Assume that the space state of X_is is finite. (The space of Y_is can be either discrete or continuous. We will keep $Y_i = y_i$ fixed and the likelihood $L(x_i, y_i)$ will be considered as a function of x_i.) Assume also that the transition probabilities T and the likelihood L are *known*. (A more complicated model usually includes some unknown parameters on which T and L depend. In our notes we only consider a simplified version of the problem.)

The general idea behind the algorithms below is recursive (sequential) computation of inference probabilities $p(x|y)$.

Forward filtering

Let

$$\alpha_i(x_i) = p(x_i, y_0, \ldots, y_i), \qquad i = 0, \ldots, k.$$

Recursive formula is the following: $\alpha_0(x_0) = s(x_0)L(x_0, y_0)$ and

$$\begin{aligned}\alpha_{i+1}(x_{i+1}) &= p(x_{i+1}, y_0, \ldots, y_{i+1}) \\ &= \sum_{x_i} p(x_i, x_{i+1}, y_0, \ldots, y_i, y_{i+1}) \\ &= \sum_{x_i} p(x_i, y_0, \ldots, y_i) T(x_i, x_{i+1}) L(x_{i+1}, y_{i+1}) \\ &= \sum_{x_i} \alpha_i(x_i) T(x_i, x_{i+1}) L(x_{i+1}, y_{i+1}).\end{aligned}$$

Let us make two important remarks. First, the filtering probability is $p(x_i|y_0, \ldots, y_i) \propto \alpha_i(x_i)$. Second, the marginal likelihood of all observations is $p(y_0, \ldots, y_k) = \sum_{x_k} \alpha_k(x_k)$. Computation of this quantity is useful for the statistical inference on unknown parameters, if they are included in the model.

Backward algorithm

Let

$$\beta_i(x_i) = p(y_{i+1}, \ldots, y_k|x_i).$$

Recursion goes backward as follows: $\beta_k(x_k) = 1$ and

$$\begin{aligned}\beta_i(x_i) &= \sum_{x_{i+1}} p(x_{i+1}, y_{i+1}, \ldots, y_k|x_i) \\ &= \sum_{x_{i+1}} p(x_{i+1}|x_i) p(y_{i+1}|x_{i+1}) p(y_{i+2}, \ldots, y_k|x_{i+1}) \\ &= \sum_{x_{i+1}} T(x_i, x_{i+1}) L(x_{i+1}, y_{i+1}) \beta_{i+1}(x_{i+1}).\end{aligned}$$

Now, it is easy to compute

$$p(x_i|y_0, \ldots, y_k) \propto p(x_i|y_0, \ldots, y_k) = \alpha_i(x_i)\beta_i(x_i).$$

Forward Filtering Backward Sampling (FFBS)

Let

$$B_i(x_{i+1}, x_i) = p(x_i|x_{i+1}, y_0, \ldots, y_k).$$

This B_i can be treated as the transition rule of the reversed (nonhomogeneous) Markov chain $X_k, \ldots, X_1, X_0|y_0, y_1, \ldots, y_k$. We have

$$\begin{aligned}B_i(x_{i+1}, x_i) &\propto p(x_i, x_{i+1}, y_0, \ldots, y_k) \\ &= \alpha_i(x_i) T(x_i, x_{i+1}) L(x_{i+1}, y_{i+1}) \beta_{i+1}(x_{i+1}) \\ &\propto \alpha_i(x_i) T(x_i, x_{i+1}).\end{aligned}$$

Recursive sampling proceeds as follows: sample $X_k \sim \alpha_k(\cdot)$ (normalized). For given $X_{i+1} = x_{i+1}$, sample $X_i \sim B_i(x_{i+1}, \cdot)$ for $i = k-1, \ldots, 1, 0$.

Viterbi algorithm

Let

$$\gamma_i(x_i) = \max_{x_0,\ldots,x_{i-1}} p(x_0, \ldots, x_{i-1}, x_i, y_0, \ldots, y_i).$$

Recursive computation is the following:

$$\begin{aligned}\gamma_{i+1}(x_{i+1}) &= \max_{x_i} \max_{x_0,\ldots,x_{i-1}} p(x_0, \ldots, x_{i-1}, x_i, y_0, \ldots, y_i) p(x_{i+1}|x_i) p(y_{i+1}|x_{i+1}) \\ &= \max_{x_i} \gamma_i(x_i) T(x_i, x_{i+1}) L(x_{i+1}|y_{i+1}).\end{aligned}$$

Maximum a posteriori (MAP) sequence $x_0^*, x_1^*, \ldots, x_k^*$ is the maximizer of

$$p(x_0, \ldots, x_k|y_0, \ldots, y_k) \propto p(x_0, \ldots, x_k, y_0, \ldots, y_k).$$

Of course, $\max p(x_0, \ldots, x_k|y_0, \ldots, y_k) = \max_{x_k} \gamma_k(x_k)$, so $x_k^* = \arg\max_{x_k} \gamma_k(x_k)$. To find the MAP path, we trace back the recursions. Let $m(x_{i+1})$ be $\arg\max_{x_i}$ when computing $\gamma_{i+1}(x_{i+1})$. Then we put $x_i^* = m(x_{i+1}^*)$ for $i = k-1, \ldots, 0$.

Chapter 5

Some supplementary theory

5.1 Intrinsic loss functions

Consider a parmetric family $\{P_\theta : \theta \in \Theta\}$ of probability distributions over $\mathcal{X}$. Let $\ell : \Theta \times \Theta \to [0, \infty[$ (so that an action consists in choosing a parameter, as in estimation problems). The main idea is that $\ell(\hat{\theta}, \theta)$ should depend only on $P_{\hat{\theta}}$ and P_θ. Put differently, the loss should be invariant with respect to re-paramerization. Such a loss function will be called *intrinsic*. Similarly, if a choice of prior is invariant with respect to re-paramerization, we speak of *intrinsic* priors.

Kullback-Leibler loss

If we have a parmetric family $\{p_\theta : \theta \in \Theta\}$ of probability densities (with respect to a common measure $\mathrm{d}x$) then an obvious candidate for an intrinsic loss function is

$$\ell(\hat{\theta}, \theta) = \mathcal{D}(p_\theta \| p_{\hat{\theta}}) = \int_{\mathcal{X}} p_\theta(x) \log \frac{p_\theta(x)}{p_{\hat{\theta}}(x)} \mathrm{d}x. \tag{5.1.1}$$

Minimization of the posterior risk corresponding to this loss function has a nice interpretation.

5.1.2 Proposition. *If ℓ is given by (5.1.1) and we observe $X_1 = x_1$ then the corresponding Bayes estimator $\hat{\theta}$ maximizes*

$$\mathbb{E}\left(\log p_{\hat{\theta}}(X_2) | X_1 = x_1\right),$$

where $X_2|\theta \sim p_\theta$ and $X_2 \perp\!\!\!\perp X_1|\theta$.

There is a clear analogy with the maximum likelihood estimation. The ML estimator maximizes

$$\log p_{\hat{\theta}}(x_1),$$

that is the likelihood of the observation. In contrast, the IBE (*Intrinsic Bayes Estimator*) in Proposition 5.1.2 maximizes the *expected* likelihood of the future observation (expectation with respect to the *predictive* distribution).

Proof of Proposition 5.1.2. The posterior risk, given x_1 is equal to

$$\begin{aligned} r_{x_1}(\hat{\theta}) &= \mathbb{E}\left(\mathcal{D}(p_\theta \| p_{\hat{\theta}})|x_1\right) = \int_\Theta \mathcal{D}(p_\theta \| p_{\hat{\theta}})\pi_{x_1}(\theta)\mathrm{d}\theta \\ &= \int_\Theta \int_{\mathcal{X}} p_\theta(x_2) \log \frac{p_\theta(x_2)}{p_{\hat{\theta}}(x_2)} \mathrm{d}x_2 \pi_{x_1}(\theta)\mathrm{d}\theta \\ &= -\int_{\mathcal{X}} \int_\Theta p_\theta(x_2)\pi_{x_1}(\theta)\mathrm{d}\theta \log p_{\hat{\theta}}(x_2)\mathrm{d}x_2 + \text{ sth independent of } \hat{\theta} \\ &= -\int_{\mathcal{X}} p(x_2|x_1) \log p_{\hat{\theta}}(x_2)\mathrm{d}x_2 + \text{ sth independent of } \hat{\theta} \\ &= -\mathbb{E}\left(\log p_{\hat{\theta}}(X_2)|x_1\right) + \text{ sth independent of } \hat{\theta}. \end{aligned}$$

□

5.1.3 REMARK. In Proposition 5.1.2, we need not assume that $X_2 =_{\mathrm{d}} X_1|\theta$. It is sufficient to define the KL loss in terms of $p_\theta(x_2)$ and not $p_\theta(x_1)$. This is clear from the proof.

Formulas for intrinsic loss in several standard exponential families are derived in Problems 4, 5 and 6 in Chapter 1. If these models are equipped with conjugate priors, then the formulas for IBEs become particularly simple. Examples are given in Problems 1 and 2 at the end of this section. A general formulation is the following.

Consider an exponential family with the natural parametrization, i.e. assume that

$$p_\theta(x) = \exp\left\{T(x)^\top \theta - \psi(\theta)\right\} h(x),$$

Recall that $\mathbb{E}(T(X)|\theta) = \nabla\psi(\theta)$. Assume that we are given a sample $x_1 \ldots, x_n$ from p_θ. It is clear that the ML $\hat{\theta}$ is the solution of the system of equations

$$\nabla\psi(\theta) = \bar{t} = \frac{1}{n}\sum_{i=1}^n T(x_i).$$

Let $\pi = \pi_{t_0,n_0}$ be a conjugate prior,

$$\pi(\theta) = c(t_0, n_0) \exp\left\{t_0^\top g(\theta) + n_0\psi(\theta)\right\},$$

and assume that (3.3.1) holds. Then for the IBE we obtain an expression which is a direct analogue of that for the ML. Since $\mathbb{E}(T(X_{n+1})|x_1,\dots,x_n) = (t_0 + n\bar{t})/(n_0+n)$, the IBE is

$$\nabla\psi(\theta) = \frac{t_0 + n\bar{t}}{n_0 + n}.$$

Intrinsic credible regions

We now proceed to an "intrinsic" approach to credible regions. The HPD regions described in Proposition 3.7.2 are "volume-optimal" but of course are not invariant with respect to re-paramerization. The Intrinsic Credible Region (ICR) is a set $C_x \subseteq \Theta$ such that

$$\mathbb{P}(\theta \in C_x|x) = \int_{C_x} \pi_x(\theta)\mathrm{d}\theta = 1-\alpha,$$

$$r_x(\hat{\theta}_1) \le r_x(\hat{\theta}_2) \text{ whenever } \hat{\theta}_1 \in C_x \text{ and } \hat{\theta}_2 \notin C_x,$$

where $r_x(\cdot)$ is the posterior risk corresponding to the KL loss (5.1.1). In other words, C_x is the sublevel set of r_x with posterior probability $1-\alpha$. The definition is simple and convincing but the computations may be difficult (a situation quite common to the whole Bayesian statistics). Usually to compute an ICR we have to resort to numerical or Monte Carlo methods. Two examples are in Problems 1 and 2.

5.2 Jeffreys priors

Given a parmetric family $\{p_\theta : \theta \in \Theta\}$ of probability densities on the observation space, we consider a prior density which is a function of the Fisher information, $\pi(\theta) = g(I(\theta))$. The question is how to choose g to ensure the invariance with respect to re-paramerization. To begin with, consider one parameter case. Let $\Theta, \Psi \subseteq \mathbb{R}$ be two open sets and introduce a new parameter $\psi = h^{-1}(\theta)$, where $h : \Theta \to \Psi$ is a '$1-1$' function (a diffeomorphism). Denote by $\tilde{I}(\psi)$ the Fisher information in the re-parametrized model and by $\tilde{\pi}(\psi)$ the density of the random variable ψ. Since

$$I(\theta) = \int \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(x)\right)^2 p_\theta(x)\mathrm{d}x$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\psi}\log p_{h(\psi)}(x) = h'(\psi)\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(x)_{|\theta=h(\psi)},$$

it follows that $\tilde{I}(\psi) = I(h(\psi))\,[h'(\psi)]^2$. On the other hand, if π is a (prior) density of θ then random variable ψ has the density $\tilde{\pi}(\psi) = \pi(h(\psi))|h'(\psi)|$. Therefore we must have

$$g(\tilde{I}(\psi)) = g\left(I(h(\psi))\,[h'(\psi)]^2\right) = g\big(I(h(\psi))\big)|h'(\psi)|.$$

Since this equation holds for every diffeomorfism h at every point ψ, it follows that $g(I) \propto \sqrt{I}$. The requirement of invariance implies that the prior is

$$\pi(\theta) \propto \sqrt{I(\theta)}. \tag{5.2.1}$$

This is the formula for *Jeffreys prior*.

5.2.2 EXAMPLE. Consider the family of Bernoulli distributions

$$p_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

From the formula for the Fisher information, $I(\theta) = \dfrac{n}{\theta(1-\theta)}$, we obtain

$$\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}.$$

Thus the Jeffreys prior is $\mathrm{Beta}(1/2, 1/2)$. Fortunately, the result is one of the conjugate priors. △

5.2.3 EXAMPLE. Consider the family of normal distributions $\mathrm{N}(\theta, \sigma^2)$ with known σ and unknown mean. The Fisher information is $I(\theta) = \sigma^{-2}$, so the Jeffreys prior should be

$$\pi(\theta) \propto 1.$$

This is an unpleasant result, because there is no probability distribution over $\Theta = \mathbb{R}$ with constant density! △

The negative result above is not the end of the story. We need a digression here.

Improper prior distributions

In statistical practice, it is quite usual to use some functions with infinite integral over the parameter space as "prior densities". They correspond to so-called "improper probability distributions" (measures of infinite total mass). The practical approach is brutal: even if $\int \pi(\theta)\mathrm{d}\theta = \infty$. We can *formally* apply the Bayes rule,

$$\pi_x(\theta) \propto p_\theta(x)\pi(\theta),$$

and it may happen that $\int \pi_x(\theta)\mathrm{d}\theta < \infty$. If we have a proper "posterior" corresponding to an improper prior, a practically oriented statistician is happy, for the inference is based on the posterior. Unfortunately, this approach is not theoretically justified and we lose the nice interpretation of Bayesian inference. (There exists a rigorous theory justifying the use of improper priors. However, this theory, created by a hungarian mathematician Rényi, is generally ignored or forgotten. Discussion on this goes beyond the scope of these notes.) For a different and more recent approach, see [7].

5.2.4 EXAMPLE (Continuation of 5.2.3). Consider the family of normal distributions with unknown mean θ. The improper Jeffreys prior for θ is the "uniform distribution" $\mathrm{U}(-\infty, \infty)$ with the density $\pi(\theta) \propto 1$. If $x_1, \ldots, x_n$ is a sample from $\mathrm{N}(\theta, \sigma^2)$ then

$$\pi_{x_1,\ldots,x_n}(\theta) \propto \exp\left(-\frac{n}{\sigma^2}(\bar{x} - \theta)^2\right).$$

Thus *a posteriori* we have $\theta \sim \mathrm{N}(\bar{x}, \sigma^2/n)$. The posterior density is equal to the normalized likelihood and the MAP (maximum a posteriori) estimator is equal to the MLE ($\hat{\theta} = \bar{x}$).

Note that the same result can be obtained if we start with the proper conjugate prior $\theta \sim \mathrm{N}(\mu, v^2)$ and take the limit with $v \to \infty$. △

5.2.5 EXAMPLE. Consider the family of normal distributions $\mathrm{N}(\mu, \sigma^2)$ with known mean μ and unknown variance. The Fisher information is $I(\sigma) = 2\sigma^{-2}$, so the Jeffreys prior is

$$\pi(\sigma) \propto \sigma^{-1}.$$

Tis is also an improper prior over $]0, \infty[$. Note that the Jeffreys prior for $\psi = \sigma^{-2}$ (the *precision* parameter) is $\tilde{\pi}(\psi) \propto \psi^{-1}$. This can be interpreted as "$\psi \sim \mathrm{Gamma}(0, 0)$ *a priori*". If $x_1, \ldots, x_n$ is a sample from $\mathrm{N}(\mu, \sigma^2)$ then *a posteriori* $\psi \sim \mathrm{Gamma}\left(n/2, \sum(x_i - \mu)^2/2\right)$. The same result can be obtained if we start with the proper conjugate prior $\psi \sim \mathrm{Gamma}(\alpha, \lambda)$ and then let $\alpha, \lambda \to 0$. △

Multivariate parameter case

The case of multivariate parameter $\theta \in \mathbb{R}^d$ can be dealt with similarly. The Jeffreys prior is given by

$$\pi(\theta) \propto \sqrt{\det I(\theta)}, \tag{5.2.6}$$

where $I(\theta)$ is the information matrix given by

$$I(\theta) = \int \left(\nabla_\theta \log p_\theta(x) \nabla_\theta^\top \log p_\theta(x)\right) p_\theta(x) \mathrm{d}x.$$

Consider a re-parametrization $\psi = h^{-1}(\theta)$ by a diffeomorphism h between two open subsets of $\mathbb{R}^d$. Denote by $\tilde{I}(\psi)$ the Fisher information in the re-parametrized model and by $\tilde{\pi}(\psi)$ the density of the random variable ψ. and

$$\nabla_\psi \log p_{h(\psi)}(x) = h'(\psi) \nabla_\theta \log p_\theta(x)_{|\theta = h(\psi)},$$

so it follows that $\tilde{I}(\psi) = h'(\psi) I(h(\psi)) h'(\psi)^\top$, where $h'(\psi)$ is a $d \times d$ nonsingular matrix. If $\pi(\theta) \propto \sqrt{\det I(\theta)}$, then $\tilde{\pi}(\psi) = \pi(h(\psi)) |\det h'(\psi)| = \sqrt{\det I(h(\psi))} |\det h'(\psi)| = \sqrt{\det h'(\psi) \det I(h(\psi)) h'(\psi)^\top} = \sqrt{\det \tilde{I}(\psi)}$. If the prior for θ obeys the Jeffreys rule, so does the prior for ψ.

5.2.7 EXAMPLE. For the family of normal distributions $\mathrm{N}(\theta, \sigma^2)$ with both parameters θ and σ unknown, the information matrix is

$$I(\theta, \sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}.$$

Therefore the Jeffreys prior is $\pi(\theta, \sigma) \propto \sigma^{-2}$ (θ is *a priori* independent of σ and has the improper distribution $\mathrm{U}(-\infty, \infty)$). △

Problems

1. Let $X_1 \perp\!\!\!\perp X_2 | \theta$ and $X_i | \theta \sim \mathrm{Bin}(n_i, \theta)$ for $i = 1, 2$. Consider the conjugate prior $\theta \sim \mathrm{Beta}(\alpha, \beta)$. Compute directly the Bayesian risk $r_{x_1}(\hat{\theta})$ corresponding to the intrinsic loss (up to a constant independent of $\hat{\theta}$, as in the proof of Proposition 5.1.2). Give a formula for the IBE. Devise a Monte Carlo algorithm for computing the ICR.

2. Let $X_1 \perp\!\!\!\perp X_2 | \theta$ and $X_i | \theta \sim \mathrm{Poiss}(t_i \theta)$ for $i = 1, 2$. Consider the conjugate prior $\theta \sim \mathrm{Gamma}(\alpha, \lambda)$. Compute directly the Bayesian risk $r_{x_1}(\hat{\theta})$ corresponding to the intrinsic loss (up to a constant). Give a formula for the IBE. Devise a Monte Carlo algorithm for computing the ICR.

3. Compute the Jeffreys prior for the family of multinomial distributions $\mathrm{Mult}(\theta_1, \ldots, \theta_d)$.

4. Compute the Jeffreys prior for the family of binomial distributions (as in Example 5.2.2), choosing log-odds ratio $\psi = \log \dfrac{\theta}{1 - \theta}$ as a new parameter.

5.3 Bayesian models without likelihoods

In this section we present a novel decision-theoretical extension of Bayesian models, proposed by Bissiri, Holmes and Walker in 2013. The main idea is the following. The basic operation in Bayesian statistics is updating the prior to the posterior:

$$\begin{array}{ccc} \pi & \longrightarrow & \pi_x \\ \text{prior} & \text{data } x & \text{posterior} \end{array}$$

This update is done via the Bayes formula, $\pi_x \propto p_\theta(x)\pi(\theta)$, and requires specification of the likelihood $p_\theta(x)$. Bissiri, Holmes and Walker propose to look at this operation from a decision-theoretical perspective and find an update rule which uses loss functions instead of likelihoods. We first describe their approach in abstract terms and later show how it generalized the standard Bayesian setup.

Suppose we have a prior distribution π on Θ, which reflects our prior beliefs o knowledge about the parameter of interest θ (which is not necessarily an identifier of any probability distribution over the space of data, $\mathcal{X}$). We specify a loss function $\rho(\theta, x)$, for $\theta \in \Theta$ and $x \in \mathcal{X}$. Upon observing $x \in \mathcal{X}$, we look for a probability distribution which is the solution of the minimization problem

$$L(\hat{\pi}|\pi, x) = \underbrace{\mathbb{E}_{\hat{\pi}}\rho(\theta, x)}_{\text{fit to data}} + \underbrace{\mathcal{D}(\hat{\pi}\|\pi)}_{\text{fidelity to prior}} \to \min_{\hat{\pi}}, \tag{5.3.1}$$

where $\hat{\pi}$ runs through all probability distributions on Θ, absolutely continuous with respect to π. The solution to (5.3.1) is named $\hat{\pi}_x$ and takes over the role of the posterior. Rather unexpectedly, $\hat{\pi}_x$ can be explicitly described and looks similarly as the Bayes formula for the posterior.

A lemma

5.3.2 Lemma. *Assume that π is a probability distribution over Θ and $\ell : \Theta \to \mathbb{R}$ is a function such that $\int \mathrm{e}^{-\ell(\theta)}\pi(\mathrm{d}\theta) < \infty$.*

Consider the problem of minimizing

$$\mathbb{E}_{\hat{\pi}}\ell(\theta) + \mathcal{D}(\hat{\pi}\|\pi) \to \min_{\hat{\pi}},$$

with respect to measures $\hat{\pi} \ll \pi$. The solution is $\hat{\pi} = \pi^*$, where

$$\pi^*(\mathrm{d}\theta) = \frac{\mathrm{e}^{-\ell(\theta)}\pi(\mathrm{d}\theta)}{\int \mathrm{e}^{-\ell(\theta')}\pi(\mathrm{d}\theta')}.$$

Proof. Without loss of generality assume that π and $\hat{\pi}$ have densities with respect to some measure. Write $\pi(\mathrm{d}\theta) = \pi(\theta)\mathrm{d}\theta$, $\hat{\pi}(\mathrm{d}\theta) = \hat{\pi}(\theta)\mathrm{d}\theta$. Let $z = \int \mathrm{e}^{-\ell(\theta)}\pi(\theta)\mathrm{d}\theta$. With this notation we have

$$\begin{aligned}
\mathbb{E}_{\hat{\pi}}\ell(\theta) + \mathcal{D}(\hat{\pi}\|\pi) &= \int \hat{\pi}(\theta)\ell(\theta)\mathrm{d}\theta + \int \hat{\pi}(\theta)\log\frac{\hat{\pi}(\theta)}{\pi(\theta)}\mathrm{d}\theta \\
&= \int \hat{\pi}(\theta)\log\frac{c\hat{\pi}(\theta)}{\mathrm{e}^{-\ell(\theta)}\pi(\theta)}\mathrm{d}\theta - \log z \\
&= \int \hat{\pi}(\theta)\log\frac{\hat{\pi}(\theta)}{\pi^*(\theta)}\mathrm{d}\theta - \log z = \mathcal{D}(\hat{\pi}\|\pi^*) - \log z,
\end{aligned}$$

and the RHS is clearly minimal for $\hat{\pi} = \pi^*$. □

The two dual versions of this simple result are noteworthy.

5.3.3 Corollary. *(i) The minimum of $\mathbb{E}_{\hat\pi}\ell(\theta)$ under the constraint $\mathcal{D}(\hat\pi\|\pi) \le c$ is attained by*

$$\pi_\beta(\mathrm{d}\theta) = \frac{\mathrm{e}^{-\beta\ell(\theta)}\pi(\mathrm{d}\theta)}{z_\beta},$$

provided that $\beta > 0$ is chosen in such a way that $\mathcal{D}(\pi_\beta\|\pi) = c$.

(ii) If $\mathbb{E}_\pi\ell(\theta) > c$ then the minimum of $\mathcal{D}(\hat\pi\|\pi)$ under the constraint $\mathbb{E}_{\hat\pi}\ell(\theta) \le c$ is attained by π_β defined above, provided that $\beta > 0$ is chosen in such a way that $\mathbb{E}_{\pi_\beta}\ell(\theta) = c$.

Indeed,

$$\mathbb{E}_{\pi_\beta}\beta\ell(\theta) + \mathcal{D}(\pi_\beta\|\pi) \le \mathbb{E}_{\hat\pi}\beta\ell(\theta) + \mathcal{D}(\hat\pi\|\pi) \le \begin{cases} \text{by constraint (i)} & \mathbb{E}_{\hat\pi}\beta\ell(\theta) + \mathcal{D}(\pi_\beta\|\pi); \\ \text{by constraint (ii)} & \mathbb{E}_{\pi_\beta}\beta\ell(\theta) + \mathcal{D}(\hat\pi\|\pi). \end{cases}$$

Note that π_β is is a Gibbs probability distribution. Part (i) of Corollary 5.3.3 is basically about maximmization of entropy under a generalized moment constraint. It is clear if we formally substitute $\pi(\mathrm{d}\theta) = \mathrm{d}\theta$, c.f. Problems 1 and 2.

Updating belief distributions

By Lemma 5.3.2, if we update π via minimizing $L(\hat\pi|\pi, x) = \mathbb{E}_{\hat\pi}\rho(\theta, x) + \mathcal{D}(\hat\pi\|\pi)$ then the minimizer is $\hat\pi = \hat\pi_x$ given by

$$\hat\pi_x(\theta) = \frac{1}{z_x}\mathrm{e}^{-\rho(\theta,x)}\pi(\theta), \tag{5.3.4}$$

where z_x is the norming constant (probability distributions are identified with densities). If $x = (x_1, \dots, x_n)$ we may consider the additive form of loss,

$$\rho(\theta, x) = \sum_{i=1}^n \rho_i(\theta, x_i), \tag{5.3.5}$$

thus treating x_is as "independent pieces of information" (this is natural e.g. if we believe that x_is are values of independent random variables X_i).

On the one hand, (5.3.4) combined with (5.3.5) leads to a (quasi-)Bayesian modification of M-estimators. Assume that all functions ρ_i in (5.3.5) are the same. M-estimator is a single value $\hat\theta_x$ which minimizes

$$\sum_{i=1}^n \rho(\hat\theta, x_i).$$

Now, instead of this single value we obtain a probability distribution $\hat{\pi}_x$ which minimizes

$$\sum_{i=1}^{n} \rho(\theta, x_i) + \int \hat{\pi}(\theta) \log \frac{\hat{\pi}(\theta)}{\pi(\theta)} \mathrm{d}\theta.$$

The spread of $\hat{\pi}_x$ represents (posterior) uncertainty about the parameter of interest.

On the other hand, (5.3.4) is a direct generalization of the standard Bayes formula. Indeed, if we put $\rho(\theta, x) = -\log p_\theta(x)$ then $\hat{\pi}_x(\theta) \propto p_\theta(x)\pi(x)$ is the classical posterior distribution. There is a bonus, however. The approach based on loss functions, in contrast to that based on the likelihood, makes sense also in the case of model misspecification (c.f. Section 1.3). We need not assume that the "true" probability distribution which generates data x belongs to the parametric family p_θ. Without this assumption, the loss function $\rho(\theta, x) = -\log p_\theta(x)$ still can be used but the meaning of the classical Bayes formula becomes unclear.

The choice of the KL-divergence in (5.3.1) is not arbitrary. It ensures the coherence of sequential updates in the following sense.

5.3.6 Proposition. *Assume that the loss function is of the form (5.3.5). The update rule defined by minimization of $L(\cdot|\pi, x)$ given by (5.3.1) has the following property:*

$$\hat{\pi}_{(x_1,x_2)} = (\widehat{\hat{\pi}_{x_1}})_{x_2},$$

where the RHS is the minimizer of $L(\cdot|\hat{\pi}_{x_1}, x_2)$, where $\hat{\pi}_{x_1}$ minimizes $L(\cdot|\pi, x_1)$ and $\hat{\pi}_{(x_1,x_2)}$ results from the direct application of (5.3.1) to the vector of data (x_1, x_2). Generalization to more than two components is straightforward.

Proposition 5.3.6 is a straightforward consequence of Lemma 5.3.2.

A sort of converse statement is also true. Consider the loss function of the form $L_g(\hat{\pi}|\pi, x) = \mathbb{E}_{\hat{\pi}}\rho(\theta, x) + \mathcal{D}_g(\hat{\pi}\|\pi)$, where

$$\mathcal{D}_g(\hat{\pi}\|\pi) = \mathbb{E}_{\hat{\pi}} g(\hat{\pi}/\pi).$$

5.3.7 Proposition. *If the update rule defined by minimization of $L_g(\cdot|\pi, x)$ has the coherence property formulated in Proposition 5.3.6, then $\mathcal{D}_g$ must be the KL-divergence $\mathcal{D}$.*

The proof is left as an excercise in Problem 3.

5.3.8 EXAMPLE. Let $\theta = (\theta_1, \theta_2, \theta_3)$. For $x \in \mathbb{R}$ define

$$\begin{aligned}\rho(\theta, x) = &\frac{3}{4}(\theta_1 - x)\mathbb{1}(x \le \theta_1) + \frac{1}{4}(x - \theta_1)\mathbb{1}(x > \theta_1) \\ &+ \frac{1}{2}(\theta_2 - x)\mathbb{1}(x \le \theta_2) + \frac{1}{2}(x - \theta_2)\mathbb{1}(x > \theta_2) \\ &+ \frac{1}{4}(\theta_3 - x)\mathbb{1}(x \le \theta_3) + \frac{3}{4}(x - \theta_3)\mathbb{1}(x > \theta_3).\end{aligned}$$

Let $\rho(\theta, x_1, \ldots, x_n) = \sum_{i=1}^n \rho(\theta, x_i)$. Of course, the M-estimator is the vector of 3 sample quartiles $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$, see (1.1.4). If we introduce a prior distribution on θ then we obtain the posterior $\hat{\pi}_{x_1,\ldots,x_n}$, which reflects some uncertainty about the sample quartiles.

If we considered only θ_2 and $\rho(\theta_2, x) = \frac{1}{2}|\theta_2 - x|$ then the result could have been interpreted as the posterior corresponding to the Laplace likelihood $p_{\theta_2}(x) = \frac{1}{2}\mathrm{e}^{-|x|}$. However, the inference on three quartiles jointly does not easily fit into the classical Bayesian framework.

Note also that if we assumed *independent* priors for θ_1, θ_2 and θ_3 then we would have three separate, independent computations, see Problem 4. However any sensible prior should satisfy the obvious constraint $\theta_1 \leq \theta_2 \leq \theta_3$, and this is incompatible with independence. $\triangle$

Problems

1. The entropy of a probability distribution (density) ν is $H(\nu) = -\int \nu(\theta) \log \nu(\theta) \mathrm{d}\theta$. Show that the maximum of $H(\nu)$ under the the constraint $\mathbb{E}_\nu \ell(\theta) \leq c$ is attained by $\nu_\beta(\theta) \propto \mathrm{e}^{\beta \ell(\theta)}$, provided that $\beta > 0$ is chosen in such a way that $\mathbb{E}_\nu \ell(\theta) = c$. *Hint:* The argument is almost identical as in Corollary 5.3.3.

2. Let $\Theta =]0, \infty[$ and consider the entropy with respect to the Lebesgue measure $\mathrm{d}\theta$. For a given c, find the probability distribution ν with maximum entropy, subject to $\mathbb{E}_\nu \theta \leq c$.

3. Prove Proposition 5.3.7. *Hint:* It is enough to consider the space $\Theta = \{\theta_1, \theta_2\}$.

4. Assume that the prior for $\theta = (\theta_1, \theta_2)$ is a product distribution $\pi = \pi^{(1)} \times \pi^{(2)}$ and $\rho(\theta_1, \theta_2, x) = \rho(\theta_1, x) + \rho(\theta_2, x)$. Show that $\hat{\pi}_x = \hat{\pi}_x^{(1)} \times \hat{\pi}_x^{(2)}$.

Chapter 6

Bayesian Asymptotics

In this chapter we consider infinite sequences of (conditionally i.i.d.) random variables, so we are going to pay more attention to technical details and be more precise. To begin with, let us recall the basic construction of a Bayesian model in a version we will work with.

The setup

As usual, we consider observation space $\mathcal{X}$ and parameter space Θ which are Polish spaces equipped with their Borel σ-fields, a transition kernel P from Θ to $\mathcal{X}$ and a probability distribution Π on Θ. Recall that

- For every $\theta \in \Theta$, $P_\theta(\cdot)$ is a probability distribution on $\mathcal{X}$.
- For every $B \in \mathfrak{X}$, function $P_\cdot(B)$ is measurable.

Note that here and in the sequel the respective σ-fields will not be explicitly mentioned, because we always assume they are Borel σ-fields in the corresponding Polish spaces. Now we consider conditionally i.i.d. sequences. It is convenient to define the canonical probability space $\Omega = (\Theta, \mathcal{X}^\infty)$ where $\mathcal{X}^\infty = \mathcal{X} \times \mathcal{X} \times \cdots$. Since Ω is a Polish space, we can equip it with its Borel σ-field (or, equivalently, with the product of Borel σ-fields on Θ and $\mathcal{X}$). The general Fubini 2.1.1 theorem together with the Kolmogorov extension theorem guarantee the existence of a measure $\mathbb{P}$ on Ω such that

$$\mathbb{P}(C \times B_1 \times \cdots \times B_n \times \mathcal{X} \times \mathcal{X} \times \cdots) = \int_C P_\theta(B_1) \cdots P_\theta(B_n) \Pi(\mathrm{d}\theta)$$

for every (Borel) $B_1, \ldots B_n \subseteq \mathcal{X}$ and every (Borel) $C \subseteq \Theta$. The product measure P_θ^∞ on $\mathcal{X}^\infty$ satisfies $P_\theta^\infty(B_1 \times \cdots \times B_n \times \mathcal{X} \times \mathcal{X} \times \cdots) = P_\theta(B_1) \cdots P_\theta(B_n)$.

Random variables $\vartheta, X_1, \ldots, X_n, \ldots$ are, of course, defined by

$$\vartheta(\theta, x_1, x_2, \ldots) = \theta, \qquad X_i(\theta, x_1, x_2, \ldots) = x_i.$$

To shorten notation we shall write $x = (x_1, x_2, \ldots) \in \mathcal{X}^\infty$ and $X = (X_1, X_2, \ldots)$.

6.1 Consistency

Doob in 1948 gave a remarkably elegant proof of a basic consistency result for Bayesian models. Roughly speaking, the posterior distribution given a large sample of conditionally i.i.d. observations converges to a measure concentrated at the 'true value' of the conditioning variable. This fact is valid under the following minimal assumption.

6.1.1 Assumption (Identifiability)**.** *If $\theta \neq \theta'$ then $P_\theta \neq P_{\theta'}$.*

In words, if we know the distribution then we know the parameter.

In the Bayesian convergence theorem we deal with almost sure weak convergence of random probability measures (the posterors). We recall the standard notation for the posterior:

$$\Pi_{x_1,\ldots,x_n}(\cdot) = \mathbb{P}(\vartheta \in \cdot | X_1 = x_1, \ldots, X_n = x_n).$$

The 'dot' stands here for a Borel subset of Θ. Let us stress that the above conditional probability is computed with respect to the probability $\mathbb{P}$ on $\Omega = (\Theta, \mathcal{X}^\infty)$. Also recall that the weak convergence of probability measures to a "Dirac's delta" can be equivalently defined as follows:

$$\Pi_n \to_{\mathrm{w}} \delta_\theta \qquad \text{iff} \qquad \Pi_n(U) \to 1 \quad \text{for every open set } U \ni \theta.$$

6.1.2 Theorem (Bayesian consistency)**.** *Assume that we have a Bayesian model defined in the previous subsection. If Assumption 6.1.1 holds then there is a set $\Theta_1 \subseteq \Theta$ with $\Pi(\Theta_1) = 1$ such that for every $\theta \in \Theta_1$,*

$$\Pi_{x_1,\ldots,x_n} \to_{\mathrm{w}} \delta_\theta \quad \textit{for } [P_\theta^\infty]\textit{-almost all sequences } (x_1, x_2, \ldots) \in \mathcal{X}^\infty.$$

Some drawback of this result is the presence of 'exceptional null set' $\Theta_0 = \Theta \setminus \Theta_1$ on which the convergence might fail. (Personally, I am no more worried about this Θ_0 than about a $[P_\theta^\infty]$-null set in $\mathcal{X}^\infty$ but some people have a different opinion.)

Doob observed that Bayesian consistency follows from the following result from the theory of martingales (theory developped to a substantial extent by himself).

6.1.3 Lemma. *If $\mathbb{E}|X| < \infty$ and we have an increasing sequence of σ-fields $\mathcal{F}_n \nearrow \mathcal{F}_\infty$ then $\mathbb{E}(X|\mathcal{F}_n) \to \mathbb{E}(X|\mathcal{F}_\infty)$ a.s.*

$\mathcal{F}_n \nearrow \mathcal{F}_\infty$ means that $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ and $\mathcal{F}_\infty = \bigvee_{n=1}^\infty \mathcal{F}_n$. The original Doob's proof was "not even a sketch". A rigorous proof is based on the following lemma.

6.1.4 Lemma (Mesurable recovery)**.** *Under Assumption 6.1.1, there exists a measurable function $h : \mathcal{X}^\infty \to \Theta$ such that*

$$h(x_1, x_2, \ldots) = \theta \quad \text{for } [P_\theta^\infty]\text{-almost all sequences } (x_1, x_2, \ldots) \in \mathcal{X}^\infty.$$

We omit the proof of this lemma. The difficult part is measurability (with respect to the corresponding Borel σ-fields in $\mathcal{X}^\infty$ and Θ).

Proof of Theorem 6.1.2. Fix a (Borel) set $U \subseteq \Theta$. Using first Lemma 6.1.3 and then Lemma 6.1.4 we have that $[\mathbb{P}]$-almost surely

$$\begin{aligned}
\Pi_{X_1,\ldots,X_n}(U) &= \mathbb{P}(\vartheta \in U | X_1, \ldots, X_n) \\
&= \mathbb{E}(\mathbb{1}(\vartheta \in U) | X_1, \ldots, X_n) \\
&\to_{n\to\infty} \mathbb{E}(\mathbb{1}(\vartheta \in U) | X_1, X_2, \ldots) \\
&= \mathbb{E}(\mathbb{1}(h(X_1, X_2, \ldots) \in U) | X_1, X_2, \ldots) \\
&= \mathbb{1}(h(X_1, X_2, \ldots) \in U) \\
&= \mathbb{1}(\vartheta \in U).
\end{aligned}$$

This $[\mathbb{P}]$-a.s. convergence can be immidiately extended from a single U to any countable collection of Us, in particular to a collection $\mathfrak{U}$ of open sets which is a basis of the topology of Θ (such a countable basis exists since Θ is Polish). Summing up, we obtained that there exists a null-set $N \subset \Omega$ such that $\mathbb{P}(N) = 0$ and for every $\omega = (\theta, x_1, x_2, \ldots) \notin N$ and for all $U \in \mathfrak{U}$,

$$\Pi_{x_1,\ldots,x_n}(U) \to \mathbb{1}(\theta \in U) \qquad (n \to \infty).$$

If we fix $\omega = (\theta, x_1, x_2, \ldots) \notin N$, the above statement implies that

$$\Pi_{x_1,\ldots,x_n}(\cdot) \to_{\mathrm{w}} \delta_\theta \qquad (n \to \infty).$$

Indeed, for every open $U_0 \ni \theta$ there exists some $U \in \mathfrak{U}$ such that $\theta \in U \subseteq U_0$ and thus $\Pi_{x_1,\ldots,x_n}(U_0) \geq \Pi_{x_1,\ldots,x_n}(U) \to 1$.

It remains to translate the obtained $[\mathbb{P}]$-a.s. convergence into P_θ^∞-a.s. convergence. Observe that $\mathbb{P}(N) = \int_\Theta P_\theta^\infty(N_\theta)\Pi(\mathrm{d}\theta)$, where $N_\theta = \{x \in \mathcal{X}^\infty : (\theta, x) \in N\}$ is the 'section of N along the x-coordinate'. □

6.2 Exchangeability

If $X_1, \ldots, X_n$ are i.i.d. conditional on ϑ then clearly their probability distribution (joint, marginalized w.r.t. ϑ) is invariant with respect to permutations of the variables. This property is named exchangeability. An *infinite* sequence of random variables $X_1, \ldots, X_n, \ldots$ is said to be **exchangeable** if for every n and every premutation $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$,

$$(X_1, \ldots, X_n) =_{\mathrm{d}} (X_{\sigma_1}, \ldots, X_{\sigma_n}).$$

The remarkable fact is that the converse is also true: every infinite exchangeable sequence of random variables is conditionally i.i.d., given some random element ϑ. This is the famous theorem due to Bruno De Finetti. We first consider a special case of binary (0-1-valued) random variables, as in the original first De Finetti's paper. We give a simple, elementary and intuitive proof in this setting. The general case with less intuitive proof is deferred to the next (optional*) section.

6.2.1 Theorem (De Finetti)**.** *If an infinite sequence $X_1, \ldots, X_n, \ldots$ of binary (0-1-valued) random variables is exchangeable then there is a probablity measure Π on the interval $[0, 1]$ such that*

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \int_0^1 \theta^k (1-\theta)^{n-k} \Pi(\mathrm{d}\theta),$$

for every sequence $(x_1, \ldots, x_n) \in \{0, 1\}^n$, where $k = \sum_{i=1}^n x_i$.

Proof. Consider fixed $(x_1, \ldots, x_n) \in \{0, 1\}^n$ with $k = \sum_{i=1}^n x_i$. Let $r > n$ and $S_r = \sum_{i=1}^r X_i$. By exchangeability,

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | S_r = m) = \frac{\binom{r-n}{m-k}}{\binom{r}{m}} = \frac{(m)_k (r-m)_{n-k}}{(r)_n}$$

$$= \frac{\frac{(m)_k}{r^k} \cdot \frac{(r-m)_{n-k}}{r^{n-k}}}{\frac{(r)_n}{r^n}}.$$

Now notice that

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \sum_m \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | S_r = m) \mathbb{P}(S_r = m),$$

and let $\theta_m = \dfrac{m}{r}$. If we write $\pi_r(\theta_m) = \mathbb{P}(S_r = m)$ then we obtain

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \sum_{\theta_m} \frac{\prod_{i=1}^{k}\left(\theta_m - \frac{i-1}{r}\right)\prod_{i=1}^{n-k}\left(1-\theta_m - \frac{i-1}{r}\right)}{\prod_{i=1}^{n}\left(1-\frac{i-1}{r}\right)}\pi_r(\theta_m)$$
$$= \int_0^1 \varphi_r^{(n,k)}(\theta)\Pi_r(\mathrm{d}\theta),$$

where Π_r is a discrete measure defined by $\Pi_r\{\theta_m\} = \pi_r(\theta_m)$ and

$$\varphi_r^{(n,k)}(\theta) \to \theta^k(1-\theta)^{n-k}$$

uniformly, as $r \to \infty$. Using Helly's theorem we infer that Π_r converges weakly along some subsequence to a probability measure, $\Pi_{r'} \to \Pi$ for $\{r'\} \subset \{r\}$. Consequently

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \int_0^1 \theta^k(1-\theta)^{n-k}\Pi(\mathrm{d}\theta).$$

□

6.2.2 REMARK. The above proof strongly suggests 'what the conditioning variable ϑ really is'. Indeed,

$$\frac{S_r}{r} \to \vartheta$$

with probability 1 for some random variable ϑ such that $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | \vartheta = \theta) = \theta^k(1-\theta)^{n-k}$. This is clear *a posteriori*, after Theorem 6.2.1 has been proved. We can first construct a probability space using 2.1.1 and then apply the Strong Law of Large Numbers.

We are going to present two examples. The first one shows how the classical Bayesian model can be constructed from the so-called Polya urn scheme via de Finetti's theorem. The second one shows an urn scheme for which de Finetti's theorem cannot be applied.

6.2.3 EXAMPLE (Polya urn). Let $X_1, \ldots, X_n, \ldots$ be binary (0-1-valued) random variables with probability distribution defined by

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \frac{[m]^k[r-m]^{n-k}}{[r]^n},$$

where $k = \sum x_i$ and $[r]^n = r(r+1)\cdots(r-n+1)$. The sequence defined in this way is clearly exchangeable. This can be illustrated as the following urn scheme. Start with an urn which contains m white balls (coded as '1') and $r-m$ black balls(coded as '0'). Then draw one ball after another. After a ball is drawn, we return this ball to the urn together with one more ball of the same colour. It is easy to see that this scheme is equivalent to the Bernoulli/Beta model, with $(X_1, \ldots, X_n | \vartheta = \theta) \sim_{\text{i.i.d.}} \mathrm{Ber}(\theta)$ and $\vartheta \sim \mathrm{Beta}(m, r-m)$. In particular, $S_r/r \to \vartheta$ almost surely, the fact not so obvious *a priori*. △

6.2.4 EXAMPLE (Sampling without replacement). Let $X_1, \ldots, X_n, \ldots$ be binary (0-1-valued) random variables with probability distribution defined by

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \frac{(m)_k (r-m)_{n-k}}{(r)_n},$$

where $k = \sum x_i$. This scheme can be regarded as drawing without replacement n balls from an urn which contains m white balls (coded as '1') and $r - m$ black balls (coded as '0'). Clearly, if $n \leq r$ then this formula defines an exchangeable probability distribution on $\{0,1\}^n$. But the sequence $(X_1, \ldots, X_n)$ cannot be embedded in an infinite exchangeable sequence $(X_1, \ldots, X_n, \ldots, X_r, X_{r+1}, \ldots)$. If it were possible then we would obtain an infinite sequence of random variables with $\mathrm{Cov}(X_i, X_j) = -m(r-m)/r^2(r-1) < 0$ and $\mathrm{Var}(X_i) = m(r-m)/r^2$. For $r+1$ exchangeable summands we would get $\mathrm{Var}(S_{r+1}) < 0$, which is impossible. Therefore the conclusion of De Finetti's theorem is false, because a conditionally i.i.d. sequence always can be infinitely prolonged. △

* General De Finetti theorem

The classical result of De Finetti can be generalized to arbitrary infinite exchangeable sequences, not necessarily with binary values. Let $X_1, \ldots, X_n, \ldots$ be an exchangeable sequence of real random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Denote by $\mathcal{E}$ the σ-field of exchangeable events in $\mathcal{F}$. It can be shown that

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n | \mathcal{E}) = \prod_{i=1}^{n} P(A_i), \tag{6.2.5}$$

where P is the transition kernel (i.e. random probability measure on $\mathcal{X}$) given by

$$P(A) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \in A)$$

almost surely (it is evident that $P(A)$ is a $\mathcal{E}$-measurable random variable for every Borel $A \subseteq \mathcal{X}$). We therefore obtain the following mixture representation of the joint probability distribution of X_is: $\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{E} \prod_{i=1}^n P(A_i) = \int_\Omega \prod_{i=1}^n P(\omega, A_i) \mathbb{P}(\mathrm{d}\omega)$. Let Π be a measure on the set $\mathcal{P}$ of all probabilty measures on $\mathbb{R}$ which is induced by mapping $P : \Omega \to \mathcal{P}$. We can rewrite the last formula as

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{\mathcal{P}} \prod_{i=1}^{n} P(A_i) \Pi(\mathrm{d}P).$$

This expression is intuitively appealing and fully analoguous to that in Theorem 6.2.1. However, we omitted many measure-theoretic details needed for rigorous formulation an a proof. A nice proof of (6.2.5), based on martingale arguments, can be found in Durrett's book **??** (Theorem 5.6.5).

Problems

1. Consider the Normal/Normal model (Example 3.2.3). Assume that $X_1, \ldots, X_n, \ldots$ are i.i.d. sample from $\mathrm{N}(\theta_0, \sigma^2)$ and $n \to \infty$. Show directly (without using 6.1.2) that almost surely we have $\pi_{x_1,\ldots,x_n} \to_{\mathrm{w}} \delta_{\theta_0}$.

   Hint: Use SLLN and compute posterior expectation and variance.

2. Consider the Poisson/Gamma model (Example 3.2.2). Assume that $X_1, \ldots, X_n, \ldots$ are i.i.d. sample from $\mathrm{Poiss}(\theta_0)$ and $n \to \infty$. Show directly that $\pi_{x_1,\ldots,x_n} \to_{\mathrm{w}} \delta_{\theta_0}$ a.s.

3. Consider the Bin/Beta model (Example 3.2.1). Assume that $X_1, \ldots, X_n, \ldots$ are i.i.d. sample from $\mathrm{Ber}(\theta_0)$ and $n \to \infty$. Show directly that $\pi_{x_1,\ldots,x_n} \to_{\mathrm{w}} \delta_{\theta_0}$ a.s.

Chapter 7

Empirical Bayes and Linear models

7.1 Introductory example

Credibility Theory is a branch of insurance mathematics. From a theoretical viewpoint, it is an example of *Empirical Bayesian* approach. In a hierarchical statistical model, we use Bayesian estimates at the lower level and, at the higher level, we use frequentist methods to estimate the prior distribution. The following example will clarify this.

7.1.1 EXAMPLE. Consider n clients of an insurance company and data describing numbers of claims for these clients during the respective periods. Data consist of n pairs (t_i, Y_i), where

> t_i is the length of the time ith client has been insured (considered as "exposure to risk"), for $i = 1, \ldots, n$.
>
> Y_i is the number of claims for ith client (in t_i years). Assume
>
> $$Y_i \sim_{\text{i.i.d.}} \text{Poiss}(t_i\theta_i),$$
>
> so that $\mathbb{E}(Y_i|\theta_i) = t_i\theta_i$ and θ_i is the mean number of claims per year fot ith client.

The goal is to estimate $\theta_1, \ldots \theta_n$. The two naive methods of estimation are the following.

- Use $\hat{\theta}_i = \bar{Y}_i = \dfrac{Y_i}{t_i}$ – individual estimator for ith client (under the assumption that the clients are "independent").
- Use $\hat{\theta} = \bar{Y} = \dfrac{\sum_i Y_i}{\sum_i t_i}$ – collective estimator (under the assumption that $\theta_1 = \cdots = \theta_k$, i.e. that the clients are "homogeneous").

Since both these approaches seem to be in some sense "extremal" and thus unacceptable, the actuaries use some "compromise", namely a weighted average of the individual and collective estimators:

$$\hat{\theta}_i = z_i \bar{Y}_i + (1 - z_i)\bar{Y}, \tag{7.1.2}$$

where z_i is the "credibility weight". Below we show how the basic credibility formula (7.1.2) can be justified, using a mixture of Bayesian approach and frequentist statistics. △

7.2 Credibility model (Bühlmann-Straub)

Consider an array of data (Y_{ij}), $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, where random variable Y_{ij} describes claims of ith client in jth year of insurance. The model includes parameters $\theta_i, \ldots, \theta_k$ (unobserved model random variables), where θ_i corresponds to ith client.

$$\begin{array}{llllllll}
\theta_1; & Y_{11}, & \ldots & Y_{1j}, & \ldots & Y_{1n_1}, & & \\
\vdots & \vdots & \ddots & \vdots & \ddots & & & \\
\theta_i; & Y_{i1}, & \ldots & Y_{ij}, & \ldots & \ldots & \ldots & Y_{in_i}, \\
\vdots & \vdots & \ddots & \vdots & \ddots & & & \\
\theta_k; & Y_{k1}, & \ldots & Y_{kj}, & \ldots & \ldots & Y_{kn_k}. &
\end{array}$$

We assume that the parameters are independently selected from a probability distribution $\pi(\cdot)$ which describes a "population of clients". Once θ_i has been selected, the claims of ith client are (conditionally) independent draws from a probability distribution $p(\cdot|\theta_i)$. Summing up,

- $\theta_1, \ldots, \theta_k \sim_{\text{i.i.d.}} \pi(\cdot)$;
- $Y_{i1}, \ldots, Y_{in_i} \sim_{\text{i.i.d.}} p(\cdot|\theta_i)$.

The joint probability distribution is

$$p((\theta_i), (y_{ij})) = \prod_{i=1}^{k} \pi(\theta_i) \prod_{j=1}^{n_i} p(y_{ij}|\theta_i).$$

The goal is estimation/prediction of $\mu(\theta_i)$, where

$$\mu(\theta) = \int y p(y|\theta) \mathrm{d}\theta,$$

so that

$$\mu(\theta_i) = \mathbb{E}(Y_{ij}|\theta_i) = \mathbb{E}(Y_{i,\text{new } j}|\theta_i).$$

Define also

$$\sigma^2(\theta) = \int (y - \mu(\theta))^2 p(y|\theta) \mathrm{d}\theta$$

and note that $\sigma^2(\theta_i) = \mathrm{Var}(Y_{ij}|\theta_i)$.

Bühlmann-Straub is one-way classification with random effects

Under the assumptions of the Bühlmann-Straub model, define

$$\begin{aligned} m &= \int \mu(\theta)\pi(\theta)\mathrm{d}\theta, \\ s^2 &= \int \sigma^2(\theta)\pi(\theta)\mathrm{d}\theta, \\ v^2 &= \int (\mu(\theta) - m)^2 \pi(\theta)\mathrm{d}\theta. \end{aligned}$$

Interpretation of these quanities is the following. We have $m = \mathbb{E}Y_{ij}$, $s^2 = \mathbb{E}\mathrm{Var}(Y_{ij}|\theta_i)$ and $v^2 = \mathrm{Var}\mathbb{E}(Y_{ij}|\theta_i)$. Thus m is the global mean (unconditional, i.e. for the whole population of clients). The variance components $s^2 + v^2 = \mathrm{Var}Y_{ij}$ correspond to two sources of variability of Y_{ij}.

Let us write

$$Y_{ij} = m + \underbrace{(\mu(\theta_i) - m)}_{=\alpha_i \text{ random effect}} + \underbrace{(Y_{ij} - \mu(\theta_i))}_{=\varepsilon_{ij} \text{ random error}}.$$

In this way we obtain a special case of *Mixed Linear Model* (MLM), namely the model of one-way classification with random effects:

$$Y_{ij} = m + \alpha_i + \varepsilon_{ij},$$

where all variables α_i and ε_{ij} are uncorrelated, $\mathbb{E}\alpha_i = 0$, $\mathbb{E}\varepsilon_{ij} = 0$, $\mathrm{Var}\alpha_i = v^2$, $\mathrm{Var}\varepsilon_{ij} = s^2$.

Computing the variances of α_i and ε_{ij} as well as covariances between all pairs of these variables is easy. For example let us check that $\mathrm{Cov}(\alpha_i, \varepsilon_{ij}) = 0$. Indeed,

$$\begin{aligned} \mathrm{Cov}(\alpha_i, \varepsilon_{ij}) &= \mathrm{Cov}(\mu(\theta_i) - m, Y_{ij} - \mu(\theta_i)) = \mathrm{Cov}(\mu(\theta_i), Y_{ij} - \mu(\theta_i)) \\ &= \mathbb{E}\mathrm{Cov}(\mu(\theta_i), Y_{ij} - \mu(\theta_i)|\theta_i) + \mathrm{Cov}(\mu(\theta_i), \mathbb{E}(Y_{ij} - \mu(\theta_i)|\theta_i) = 0 + 0. \end{aligned}$$

The first term is 0 because $\mathrm{Cov}(\mu(\theta_i), Y_{ij} - \mu(\theta_i)|\theta_i) = 0$ (covariance of a constant with anything is zero); the second term is 0 because $\mathbb{E}(Y_{ij} - \mu(\theta_i)|\theta_i) = 0$. (Note that we have considered two variables in the same ith row of the array; variables from different rows are uncorrelated because they are independent.)

Note that in our new notation the variables to be estimated/predicted become $\mu(\theta_i) = m + \alpha_i$.

BLP, BLUP and EBLUP in the Bühlmann-Straub model

Consider the mixed linear model:

$$Y_{ij} = m + \alpha_i + \varepsilon_{ij}. \tag{7.2.1}$$

Recall that m is deterministic, α_i, ε_{ij} are uncorrelated, $\mathbb{E}\alpha_i = 0$, $\mathbb{E}\varepsilon_{ij} = 0$, $\mathrm{Var}\alpha_i = v^2$, $\mathrm{Var}\varepsilon_{ij} = s^2$. We are to estimate/predict $\mu_i = m + \alpha_i$, for all i. Let us focus on μ_i for some fixed row i. The criterion is minimization of the Mean Square Error (MSE)

$$\mathrm{MSE} = \mathbb{E}(\hat{\mu}_i - \mu_i)^2.$$

We restrict attention to linear predictors, i.e. variables of the form $\hat{\mu}_i = a_0 + \sum_{l=1}^{k}\sum_{j=1}^{n_l} a_{lj}Y_{lj}$. A simple argument based on symmetry shows that we need only consider predictors with equal coefficients a_{lj} in every row l. Equivalently, we can write

$$\hat{\mu}_i = a_0 + \sum_{l=1}^{k} a_l \bar{Y}_l.$$

BLP (*Best Linear Predictor*): Assume that m, s^2 and v^2 are known. BLP is the linear predictor which minimizes the MSE. In the model (7.2.1) the BLP is given by the following formula:

$$\hat{\mu}_i = z_i\bar{Y}_i + (1 - z_i)m, \tag{7.2.2}$$

where

$$\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}; \qquad z_i = \frac{n_i v^2}{n_i v^2 + s^2}.$$

Note that the BLP of μ_i depends only on the variables in ith row of data (individual data). We say that z_i is the *credibility coefficient*.

Derivation of (7.2.2). Note that $\hat{\mu}_i$ depends only on $\bar{Y}_i$ and not on Y_{lj} for $l \neq i$, because $\mathrm{Cov}(\mu_i, Y_{lj}) = 0$. Therefore we are to minimize $\mathrm{MSE} = \mathbb{E}(\mu_i - a_0 - a\bar{Y}_i)^2$. Minimizing with respect to a_0 yields $a_0 = \mathbb{E}\mu_i - a\mathbb{E}\bar{Y}_i$ and

$$\mathrm{MSE} = \mathrm{Var}(\mu_i - a_0 - a\bar{Y}_i) = \mathrm{Var}(\mu_i) + 2a\mathrm{Cov}(\mu_i, \bar{Y}_i) + a^2\mathrm{Var}(\bar{Y}_i),$$

The minimum MSE is obtained for $a = \mathrm{Cov}(\mu_i, \bar{Y}_i)/\mathrm{Var}(\bar{Y}_i)$. Since $\mathbb{E}\mu_i = \mathbb{E}\bar{Y}_i = m$, $\mathrm{Var}\bar{Y}_i = \mathrm{Var}(\alpha_i + \bar{\varepsilon}_i) = \mathrm{Var}(\alpha_i) + \mathrm{Var}\bar{\varepsilon}_i = v^2 + s^2/n_i$ and $\mathrm{Cov}(\mu_i, \bar{Y}_i) = \mathrm{Cov}(\alpha_i + \bar{\varepsilon}_i, \alpha_i) = \mathrm{Var}(\alpha_i) = v^2$, we obtain $a_0^* = (1 - z_i)m$ and $a^* = z_i$. □

BLUP (*Best Linear Unbiased Predictor*): Assume that m is unknown, s^2 and v^2 are known. BLUP is the linear predictor which minimizes the MSE subject to the unbiasedness constraint: $\mathbb{E}\hat{\mu}_i = \mathbb{E}\mu_i$ for all values of m. In the model (7.2.1) the BLUP is

$$\hat{\mu}_i = z_i\bar{Y}_i + (1 - z_i)\bar{Y}, \tag{7.2.3}$$

where z_i is the the credibility coefficient defined above and

$$\bar{Y} = \frac{\sum\limits_{i=1}^{k} z_i\bar{Y}_i}{\sum\limits_{i=1}^{k} z_i}.$$

Derivation of (7.2.3). Consider $\hat{\mu}_i = a_0 + \sum_{l=1}^{k} a_l\bar{Y}_l$. Since $\mathbb{E}\mu_i = \mathbb{E}\bar{Y}_l = m$, the unbiasedness condition implies $a_0 = 0$ and $\sum_{l=1}^{l} a_l = 1$. Taking this condition into account, we obtain

$$\mathrm{MSE} = \mathrm{Var}\left(\mu_i - \sum_{l=1}^{k} a_l\bar{Y}_l\right) = \mathrm{Var}(\mu_i) - 2\sum_{l=1}^{k} a_l\mathrm{Cov}(\mu_i, \bar{Y}_l) + \sum_{l=1}^{k} a_l^2\mathrm{Var}(\bar{Y}_l).$$

To find the minimum we use the method of Lagrange multipliers. The Lagrange function is $\mathcal{L} = \mathrm{MSE} - 2\lambda\sum_{l=1}^{k} a_l$. Taking the derivatives we obtain the following equations:

$$\begin{aligned}
\frac{1}{2}\frac{\partial}{\partial a_l}\mathcal{L} &= a_l\left(v^2 + s^2/n_l\right) - \lambda \quad \text{for } l \neq i,\\
\frac{1}{2}\frac{\partial}{\partial a_i}\mathcal{L} &= a_i\left(v^2 + s^2/n_i\right) - \lambda - v^2.
\end{aligned}$$

Therefore, using the formula $z_l = 1/(v^2 + s^2/n_l)$ we get

$$\begin{aligned}
a_l &= \frac{\lambda}{v^2 + s^2/n_l} = \frac{z_l}{v^2}\lambda \quad \text{for } l \neq i,\\
a_i &= \frac{\lambda}{v^2 + s^2/n_i} + \frac{v^2}{v^2 + s^2/n_i} = \frac{z_i}{v^2}\lambda + z_i.
\end{aligned}$$

It remains to compute λ using the constraint equation. With the notation $z_\bullet = \sum_{l=1}^{k} z_l$, this equation can be rewritten as $1 = \sum_{l=1}^{k} a_l = (z_\bullet/v^2)\lambda + z_i$. Consequently $\lambda = (1 - z_i)v^2/z_\bullet$ and finally

$$\begin{aligned}
a_l &= (1 - z_i)\frac{z_l}{z_\bullet} \quad \text{for } l \neq i,\\
a_i &= (1 - z_i)\frac{z_i}{z_\bullet} + z_i,
\end{aligned}$$

what was to be shown. □

A minor modification of the above derivation leads to the conclusion that BLUE, the *Best Linear Unbiased Estimator* of m is $\hat{m} = \bar{Y} = \sum_i (z_i/z_\bullet)\bar{Y}_i$. This is the weighted average with weights *proportional to the credibility coefficients*.

EBLUP (*Empirical* BLUP): If the variance components s^2 and v^2 are unknown, we have to construct estimates $\hat{s}^2$ and $\hat{v}^2$. Then we replace the credibility coefficients $z_i = n_i v^2/(n_i v^2 + s^2)$ by $\hat{z}_i = n_i \hat{v}^2/(n_i \hat{v}^2 + \hat{s}^2)$ and use the formulas for BLUP. The predictor obtained in this way is called EBLUP. (Strictly speaking, EBLUP is neither linear nor unbiased and is not necessarily the "best".)

Estimation of the variance components

An (unbiased) estimator of s^2 is standard:

$$\tilde{s}^2 = \frac{1}{\sum_{i=1}^k n_i - k} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2. \tag{7.2.4}$$

Estimation of v^2 is difficult. Below we present just one of several methods. Begin with the observation that

$$\begin{aligned}
\mathbb{E} \sum_i z_i (\bar{Y}_i - \bar{Y})^2 &= \sum_i z_i \mathrm{Var}(\bar{Y}_i - \bar{Y}) \\
&= \sum_i z_i \left(\mathrm{Var}\bar{Y}_i + \mathrm{Var}\bar{Y} - 2\mathrm{Cov}(Y_i, \bar{Y}) \right) \\
&= \sum_i z_i \left(v^2 + \frac{s^2}{n_i} + \frac{v^2}{z_\bullet} - 2\frac{v^2}{z_\bullet} \right) = \sum_i z_i \left(\frac{v^2}{z_i} - \frac{v^2}{z_\bullet} \right) \\
&= (k-1)v^2,
\end{aligned}$$

because $\mathrm{Var}\bar{Y} = \sum_i (z_i^2/z_\bullet^2)(v^2 + s^2/n_i) = \sum_i (z_i^2/z_\bullet^2)(v^2/z_i) = v^2/z_\bullet$ and $\mathrm{Cov}(\bar{Y}_i, \bar{Y}) = (z_i/z_\bullet)(v^2 + s^2/n_i) = v^2/z_\bullet$. The formula

$$\tilde{v}^2 = \frac{1}{k-1} \sum_i z_i (\bar{Y}_i - \bar{Y})^2 \tag{7.2.5}$$

would give an unbiased estimator of v^2 if it were not for the fact that $\tilde{v}^2$ is *not an estimator* at all! Unfortunately $\tilde{v}^2$ depends on the credibility coefficients z_i, which in turn depend on v^2. However, we obtain a reasonable (but biased) estimator solving the system of equations (7.2.5) together with the formulas for z_is (e.g. iterating these formulas to approximate a fixed point). Of course, we also have to use estimates of s^2 given by (7.2.4).

More on the estimation of the variance compontents is in the monograph Searle, Casella & McCulloch, *Variance Components* (2006).

Linear prediction

Two simple general rules can be extracted from the derivations in the previous subsections.

7.2.6 Lemma. *Let Y be a one-dimensional random variable, $\mathbb{E}U^2 < \infty$ and $\mathbb{E}Y^2 < \infty$. Consider (nonhomogeneous) linear predictors $\hat{U} = a_0 + aY$. The predictor which minimizes the MSE is $U^* = a_0^* + a^*Y$, where*

$$a^* = \frac{\mathrm{Cov}(Y,U)}{\mathrm{Var}(Y)}, \qquad a_0^* = \mathbb{E}U - a^*\mathbb{E}Y.$$

This U^* is called the Best Linear Predictor (BLP).

7.2.7 Lemma. *Let $Y_1, \ldots, Y_k$ be independent (or uncorrelated) random variables with probability distribution depending on a location parameter m. Assume $\mathbb{E}U = \mathbb{E}Y_i = m$ and all variances/covariances do not depend on m. The linear estimator of m which is unbiased and minimizes the MSE is $m^* = \sum_{i=1}^k a_i^* Y_i$, where*

$$a_i^* \propto \frac{1}{\mathrm{Var}(Y_i)}, \qquad \sum_{i=1}^k a_i^* = 1.$$

This m^* is called the Best Linear Unbiased Estimator (BLUE).

Some bibliographical notes

Two excellent and comprehensive monographs/textbooks on Bayesian statistics and decision theory are [2] and [10]. An older but still interesting exposition is [4]. This monograph is also available in Polish translation. Shorter exposition of basic Bayesian concepts can be found in the notes [1] and [11].

When preparing these notes, I have used only a few original papers. Section 1.2 (asymptotics of convex M-estimators) is based on [9]. Proposition 2.2.9 (Bayesian versus frequentist sufficiency) is repeated after [8]. Section 5.1 is based on [1], but in a modified form. I have changed the intrinsic loss from $\mathcal{D}(\hat{\theta}\|\theta) \vee \mathcal{D}(\theta\|\hat{\theta})$ to just $\mathcal{D}(\hat{\theta}\|\theta)$. Section 5.3 (loss-based update of prior to posterior) is taken from [3]. A thorough and rigorous teatment of the Doob's theorem can be found in 6.1.2

In these notes I assumed some basic knowledge of measure-theoretic probability. There are many good textbooks and monographs on this. My favourite is [5], because it is rasonably simple, concise (and readily accessible).

Bibliography

[1] José M. Bernardo, Chapter *Bayesian Statistics* published in the volume Probability and Statistics (R. Viertl, ed) of the Encyclopedia of Life Support Systems (EOLSS). Oxford, UK: UNESCO, 2003.

[2] José M. Bernardo and Adrian F. M. Smith, *Bayesian Theory*, John Wiley and Sons, 1994.

[3] P. G. Bissiri, C. C. Holmes and S. G. Walker, A General Framework for Updating Belief Distributions, *Journal of the Royal Statistical Siciety (B) 78, 5, 1103–1130*, 2016.

[4] Morris H. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill 1970.

[5] Rick Durrett, *Probability: Theory and Examples*, Cambridge University Press 2010 (4th edition).

[6] J. W. Miller *A detailed treatment of Doob's theorem*, arXiv:1801.03122, 2018.

[7] B. H. Lindkvist, G. Taraldsen, On the proper treatment of improper distributions, J*ournal of Statistical Planning and Inference* 195, 93–104, 2018.

[8] K. Furmańczyk and W. Niemiro, Sufficiency in Bayesian models, *Applicationes Math. 25, 1, 113–120*, 1998.

[9] W. Niemiro, M-estimators defined by convex minimization, *Ann. Statist. 20, 3, 1514-1533*, 1992.

[10] Christian P. Robert, *The Bayesian Choice. A decision-theoretic motivation*, Springer-Verlag 1994.

[11] B. Walsh, Introduction to Bayesian Analysis, Lecture Notes for EEB 596z, 2002.