Computing the derivative of $R_{\text{emp}}[f]$ with respect to $\boldsymbol{\alpha}$ and defining $F_{ij} := f_i(x_j)$, we can see that the minimum of (3.16) is achieved if

$$F^\top \mathbf{y} = F^\top F \boldsymbol{\alpha}. \tag{3.17}$$

A sufficient condition for (3.17) is $\boldsymbol{\alpha} = \left(F^\top F\right)^{-1} F^\top \mathbf{y}$ where $\left(F^\top F\right)^{-1}$ denotes the (pseudo-)inverse of the matrix.

Condition of a Matrix

If $F^\top F$ has a bad condition number (i.e. the quotient between the largest and the smallest eigenvalue of $F^\top F$ is large), it is numerically difficult [423, 530] to solve (3.17) for $\boldsymbol{\alpha}$. Furthermore, if $n > m$, i.e. if we have more basis functions f_i than training patterns x_i, there will exist a subspace of solutions with dimension at least $n - m$, satisfying (3.17). This is undesirable both practically (speed of computation) and theoretically (we would have to deal with a whole class of solutions rather than a single one).

One might also expect that if $\mathcal{F}$ is too rich, the discrepancy between $R_{\text{emp}}[f]$ and $R[f]$ could be large. For instance, if F is an $m \times m$ matrix of full rank, $\mathcal{F}$ contains an f that predicts all target values y_i correctly on the training data. Nevertheless, we cannot expect that we will also obtain zero prediction error on unseen points. Chapter 4 will show how these problems can be overcome by adding a so-called regularization term to $R_{\text{emp}}[f]$.

3.3 A Statistical Perspective

Given a particular pattern $\tilde{x}$, we may want to ask what risk we can expect for it, and with which *probability* the corresponding loss is going to occur. In other words, instead of (or in addition to) $\mathbf{E}\left[c(\tilde{x}, y, f(\tilde{x})\right]$ for a fixed $\tilde{x}$, we may want to know the distribution of y given $\tilde{x}$, i.e., $\mathrm{P}(y|\tilde{x})$.

(Bayesian) statistics (see [338, 432, 49, 43] and also Chapter 16) often attempt to estimate the density corresponding to the random variables (x, y), and in some cases, we may really *need* information about $p(x, y)$ to arrive at the desired conclusions given the training data (e.g., medical diagnosis). However, we always have to keep in mind that if we model the density p first, and subsequently, based on this approximation, compute a minimizer of the expected risk, we will have to make two approximations. This could lead to inferior or at least not easily predictable results. Therefore, wherever possible, we should avoid solving a more general problem, since additional approximation steps might only make the estimates worse [561].

3.3.1 Maximum Likelihood Estimation

All this said, we still may want to compute the conditional density $p(y|x)$. For this purpose we need to model how y is generated, based on some underlying dependency $f(x)$; thus, we specify the functional form of $p(y|x, f(x))$ and maximize the expression with respect to f. This will provide us with the function f that is *most likely* to have generated the data.

Definition 3.5 (Likelihood) *The likelihood of a sample* $(x_1, y_1), \ldots (x_m, y_m)$ *given an underlying functional dependency* f *is given by*

$$p(\{x_1, \ldots, x_m\}, \{y_1, \ldots, y_m\}|f) = \prod_{i=1}^{m} p(x_i, y_i|f) = \prod_{i=1}^{m} p(y_i|x_i, f)p(x_i) \tag{3.18}$$

Strictly speaking the likelihood only depends on the values $f(x_1), \ldots, f(x_m)$ rather than being a functional of f itself. To keep the notation simple, however, we write $p(\{x_1, \ldots, x_m\}, \{y_1, \ldots, y_m\}|f)$ instead of the more heavyweight expression $p(\{x_1, \ldots, x_m\}, \{y_1, \ldots, y_m\}|\{f(x_1), \ldots, f(x_m)\})$.

For practical reasons, we convert products into sums by taking the negative logarithm of $\mathrm{P}(\{x_1, \ldots, x_m\}, \{y_1, \ldots, y_m\}|f)$, an expression which is then conveniently minimized. Furthermore, we may drop the $p(x_i)$ from (3.18), since they do not depend on f. Thus maximization of (3.18) is equivalent to minimization of the *Log-Likelihood*

Log-Likelihood

$$\mathcal{L}[f] := \sum_{i=1}^{m} -\ln p(y_i|x_i, f). \tag{3.19}$$

Regression

Remark 3.6 (Regression Loss Functions) *Minimization of* $\mathcal{L}[f]$ *and of* $R_{\text{emp}}[f]$ *coincide if the loss function* c *is chosen according to*

$$c(x, y, f(x)) = -\ln p(y|x, f). \tag{3.20}$$

Assuming that the target values y *were generated by an underlying functional dependency* f *plus additive noise* ξ *with density* p_ξ*, i.e.* $y_i = f_{\text{true}}(x_i) + \xi_i$*, we obtain*

$$c(x, y, f(x)) = -\ln p_\xi(y - f(x)). \tag{3.21}$$

Things are slightly different in classification. Since all we are interested in is the probability that pattern x has label 1 or -1 (assuming binary classification), we can transform the problem into one of estimating the logarithm of the probability that a pattern assumes its correct label.

Classification

Remark 3.7 (Classification Loss Functions) *We have a finite set of labels, which allows us to model* $\mathrm{P}(y|f(x))$ *directly, instead of modelling a density. In the binary classification case (classes* 1 *and* -1*) this problem becomes particularly easy, since all we have to do is assume functional dependency underlying* $\mathrm{P}(1|f(x))$*: this immediately gives us* $\mathrm{P}(-1|f(x)) = 1 - \mathrm{P}(1|f(x))$. *The link to loss functions is established via*

$$c(x, y, f(x)) = -\ln \mathrm{P}(y|f(x)). \tag{3.22}$$

The same result can be obtained by minimizing the cross entropy[6] *between the classifica-*

[6] In the case of discrete variables the cross entropy between two distributions P and Q is defined as $\sum_i \mathrm{P}(i) \ln \mathrm{Q}(i)$.

Table 3.1 Common loss functions and corresponding density models according to Remark 3.6. As a shorthand we use $\tilde{c}(f(x)-y) := c(x, y, f(x))$.

	loss function $\tilde{c}(\xi)$	density model $p(\xi)$
ε-insensitive	$\lvert\xi\rvert_\varepsilon$	$\frac{1}{2(1+\varepsilon)}\exp(-\lvert\xi\rvert_\varepsilon)$
Laplacian	$\lvert\xi\rvert$	$\frac{1}{2}\exp(-\lvert\xi\rvert)$
Gaussian	$\frac{1}{2}\xi^2$	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{\xi^2}{2})$
Huber's robust loss	$\begin{cases} \frac{1}{2\sigma}(\xi)^2 & \text{if } \lvert\xi\rvert \leq \sigma \\ \lvert\xi\rvert - \frac{\sigma}{2} & \text{otherwise} \end{cases}$	$\propto \begin{cases} \exp(-\frac{\xi^2}{2\sigma}) & \text{if } \lvert\xi\rvert \leq \sigma \\ \exp(\frac{\sigma}{2} - \lvert\xi\rvert) & \text{otherwise} \end{cases}$
Polynomial	$\frac{1}{d}\lvert\xi\rvert^d$	$\frac{d}{2\Gamma(1/d)}\exp(-\lvert\xi\rvert^d)$
Piecewise polynomial	$\begin{cases} \frac{1}{d\sigma^{d-1}}\lvert\xi\rvert^d & \text{if } \lvert\xi\rvert \leq \sigma \\ \lvert\xi\rvert - \sigma\frac{d-1}{d} & \text{otherwise} \end{cases}$	$\propto \begin{cases} \exp(-\frac{\lvert\xi\rvert^d}{d\sigma^{d-1}}) & \text{if } \lvert\xi\rvert\sigma \\ \exp(\sigma\frac{d-1}{d} - \lvert\xi\rvert) & \text{otherwise} \end{cases}$

tion labels y_i and the probabilities $p(y|f(x))$, as is typically done in a generalized linear models context (see e.g., [355, 232, 163]). For binary classification (with $y \in \{\pm 1\}$) we obtain

$$c(x, y, f(x)) = \frac{1+y}{2}\ln \mathrm{P}(y = 1|f(x)) + \frac{1-y}{2}\ln \mathrm{P}(y = -1|f(x)). \tag{3.23}$$

When substituting the actual values for y into (3.23), this reduces to (3.22).

At this point we have a choice in modelling $\mathrm{P}(y = 1|f(x))$ to suit our needs. Possible models include the logistic transfer function, the probit model, the inverse complementary log-log model. See Section 16.3.5 for a more detailed discussion of the choice of such *link functions*. Below we explain connections in some more detail for the logistic link function.

For a logistic model, where $\mathrm{P}(y = \pm 1|x, f) \propto \exp(\pm\frac{1}{2}f(x))$, we obtain after normalization

$$\mathrm{P}(y = 1|x, f) := \frac{\exp(f(x))}{1 + \exp(f(x))} \tag{3.24}$$

and consequently $-\ln \mathrm{P}(y = 1|x, f) = \ln(1 + \exp(-f(x)))$. We thus recover (3.5) as the loss function for classification. Choices other than (3.24) for a map $\mathbb{R} \to [0, 1]$ will lead to further loss functions for classification. See [579, 179, 596] and Section 16.1.1 for more details on this subject.

It is important to note that not every loss function used in classification corresponds to such a density model (recall that in this case, the probabilities have to add up to 1 for any value of $f(x)$). In fact, one of the most popular loss functions, the soft margin loss (3.3), does not enjoy this property. A discussion of these issues can be found in [521].

Examples

Table 3.1 summarizes common loss functions and the corresponding density models as defined by (3.21), some of which were already presented in Section 3.1. It is an exhaustive list of the loss functions that will be used in this book for regression. Figure 3.2 contains graphs of the functions.

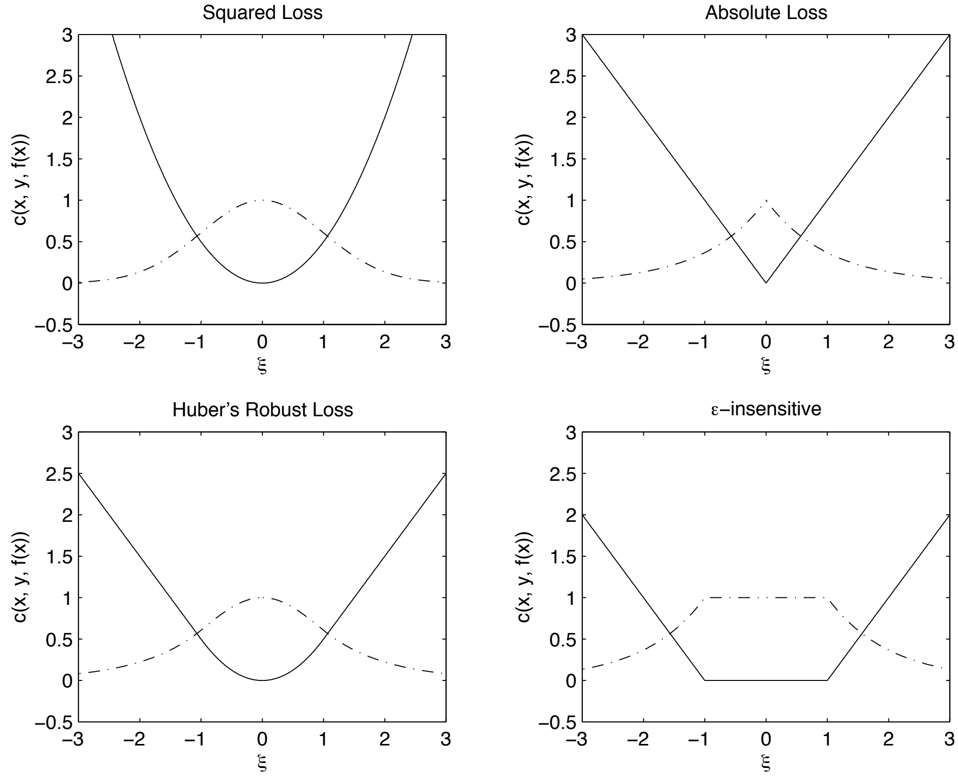

Figure 3.2 Graphs of loss functions and corresponding density models. upper left: Gaussian, upper right: Laplacian, lower left: Huber's robust, lower right: ε-insensitive.

Practical Considerations

We conclude with a few cautionary remarks. The loss function resulting from a maximum likelihood reasoning might be non-convex. This might spell trouble when we try to find an efficient solution of the corresponding minimization problem. Moreover, we made a very strong assumption by claiming to know $\mathrm{P}(y|x, f)$ explicitly, which was necessary in order to evaluate (3.20).

Finally, the solution we obtain by minimizing the log-likelihood depends on the class of functions $\mathcal{F}$. So we are in no better situation than by minimizing $R_{\text{emp}}[f]$, albeit with the additional constraint, that the loss functions $c(x, y, f(x))$ must correspond to a probability density.

3.3.2 Efficiency

The above reasoning could mislead us into thinking that the choice of loss function is rather arbitrary, and that there exists no good means of assessing the performance of an estimator. In the present section we will develop tools which can be used to compare estimators that are derived from different loss functions. For this purpose we need to introduce additional statistical concepts which deal with the efficiency of an estimator. Roughly speaking, these give an indication of how "noisy" an estimator is with respect to a reference estimator.

We begin by formalizing the concept of an estimator. Denote by $\mathrm{P}(y|\theta)$ a distribution of y depending (amongst other variables) on the parameters θ, and by $Y = \{y_1, \ldots, y_m\}$ an m-sample drawn iid from $\mathrm{P}(y|\theta)$. Note that the use of the symbol y bears no relation to the y_i that are outputs of some functional dependency (cf. Chapter 1). We employ this symbol because some of the results to be derived will later be applied to the outputs of SV regression.

Estimator

Next, we introduce the *estimator* $\hat{\theta}(Y)$ of the parameters θ, based on Y. For instance, $\mathrm{P}(y|\theta)$ could be a Gaussian with fixed variance and mean θ, and $\hat{\theta}(Y)$ could be the estimator $(1/m)\sum_{i=1}^{m} y_i$.

To avoid cumbersome notation, we use the shorthand

$$\mathbf{E}_\theta\left[\xi(y)\right] := \mathbf{E}_{\mathrm{P}(y|\theta)}\left[\xi(y)\right] = \int \xi(y) d\mathrm{P}(y|\theta), \tag{3.25}$$

to express expectations of a random variable $\xi(y)$ with respect to $\mathrm{P}(y|\theta)$. One criterion that we might impose on an estimator is that it be unbiased, i.e., that on average, it tells us the correct value of the parameter it attempts to estimate.

Definition 3.8 (Unbiased Estimator) *An unbiased estimator $\hat{\theta}(Y)$ of the parameters θ in $\mathrm{P}(y|\theta)$ satisfies*

$$\mathbf{E}_\theta\left[\hat{\theta}(Y)\right] = \theta. \tag{3.26}$$

In this section, we will focus on unbiased estimators. In general, however, the estimators we are dealing with in this book will not be unbiased. In fact, they will have a bias towards 'simple', low-complexity functions. Properties of such estimators are more difficult to deal with, which is why, for the sake of simplicity, we restrict ourselves to the unbiased case in this section. Note, however, that "biasedness" is not a bad property by itself. On the contrary, there exist cases as the one described by James and Stein [262] where biased estimators consistently outperform unbiased estimators in the finite sample size setting, both in terms of variance and prediction error.

A possible way to compare unbiased estimators is to compute their variance. Other quantities such as moments of higher order or maximum deviation properties would be valid criteria as well, yet for historical and practical reasons the variance has become a standard tool to benchmark estimators. The Fisher information matrix is crucial for this purpose since it will tell us via the Cramér-Rao bound (Theorem 3.11) the minimal possible variance for an unbiased estimator. The idea is that the smaller the variance, the lower (typically) the probability that $\hat{\theta}(Y)$ will deviate from θ by a large amount. Therefore, we can use the variance as a possible one number summary to compare different estimators.

Definition 3.9 (Score Function, Fisher Information, Covariance) *Assume there exists a density $p(y|\theta)$ for the distribution $\mathrm{P}(y|\theta)$ such that $\ln p(y|\theta)$ is differentiable with*

Score Function

respect to θ. The score $V_\theta(Y)$ *of* $\mathrm{P}(y|\theta)$ *is a random variable defined by*[7]

$$V_\theta(Y) := \partial_\theta \ln p(Y|\theta) = \partial_\theta \sum_{i=1}^{m} \ln p(y_i|\theta) = \sum_{i=1}^{m} \frac{\partial_\theta p(y_i|\theta)}{p(y_i|\theta)}. \tag{3.27}$$

This score tells us how much the likelihood of the data depends on the different components of θ, and thus, in the maximum likelihood procedure, how much the data affect the choice of θ. The covariance of $V_\theta(Y)$ is called the Fisher information matrix *I. It is given by*

Fisher Information

$$I_{ij} := \mathbf{E}_\theta \left[\partial_{\theta_i} \ln p(Y|\theta) \cdot \partial_{\theta_j} \ln p(Y|\theta) \right]. \tag{3.28}$$

Covariance

and the covariance matrix *B of the estimator $\hat{\theta}(Y)$ is defined by*

$$B_{ij} := \mathbf{E}_\theta \left[\left(\hat{\theta}_i - \mathbf{E}_\theta \left[\hat{\theta}_i \right] \right) \left(\hat{\theta}_j - \mathbf{E}_\theta \left[\hat{\theta}_j \right] \right) \right]. \tag{3.29}$$

The covariance matrix B tells us the amount of variation of the estimator. It can therefore be used (e.g., by Chebychev's inequality) to bound the probability that $\hat{\theta}(Y)$ deviates from θ by more than a certain amount.

Remark 3.10 (Expected Value of Fisher Score) *One can check that the expected value of $V_\theta(Y)$ is 0 since*

$$\mathbf{E}_\theta \left[V_\theta(Y) \right] = \int p(Y|\theta) \partial_\theta \ln p(Y|\theta) dY = \partial_\theta \int p(Y|\theta) dY = \partial_\theta 1 = 0. \tag{3.30}$$

Average Fisher Score Vanishes

In other words, the contribution of Y to the adjustment of θ averages to 0 over all possible Y, drawn according to $\mathrm{P}(Y|\theta)$. Equivalently we could say that the average likelihood for Y drawn according to $\mathrm{P}(Y|\theta)$ is extremal, provided we choose θ: the derivative of the expected likelihood of the data $\mathbf{E}_\theta \left[\ln \mathrm{P}(Y|\theta) \right]$ with respect to θ vanishes. This is also what we expect, namely that the "proper" distribution is on average the one with the highest likelihood.

The following theorem gives a lower bound on the variance of an estimator, i.e. B is found in terms of the Fisher information I. This is useful to determine how well a given estimator performs with respect to the one with the lowest possible variance.

Theorem 3.11 (Cramér and Rao [425]) *Any unbiased estimator $\hat{\theta}(Y)$ satisfies*

$$\det IB \geq 1. \tag{3.31}$$

Proof We prove (3.31) for the scalar case. The extension to matrices is left as an exercise (see Problem 3.10). Using the Cauchy-Schwarz inequality, we obtain

$$\left(\mathbf{E}_\theta \left[\left(V_\theta(Y) - \mathbf{E}_\theta \left[V_\theta(Y) \right] \right) \left(\hat{\theta}(Y) - \mathbf{E}_\theta \left[\hat{\theta}(Y) \right] \right) \right] \right)^2 \tag{3.32}$$

$$\leq \mathbf{E}_\theta \left[\left(V_\theta(Y) - \mathbf{E}_\theta \left[V_\theta(Y) \right] \right)^2 \right] \mathbf{E}_\theta \left[\left(\hat{\theta}(Y) - \mathbf{E}_\theta \left[\hat{\theta}(Y) \right] \right)^2 \right] = IB. \tag{3.33}$$

[7] Recall that $\partial_\theta p(Y|\theta)$ is the gradient of $p(Y|\theta)$ with respect to the parameters $\theta_1, \ldots, \theta_n$.

At the same time, $\mathbf{E}_\theta\left[V_\theta(Y)\right] = 0$ implies that

$$\left(\mathbf{E}_\theta\left[\left(V_\theta(Y) - \mathbf{E}_\theta\left[V_\theta(Y)\right]\right)\left(\hat{\theta}(Y) - \mathbf{E}_\theta\left[\hat{\theta}(Y)\right]\right)\right]\right)^2 \tag{3.34}$$

$$= \mathbf{E}_\theta\left[V_\theta(Y)\hat{\theta}(Y)\right]^2 \tag{3.35}$$

$$= \left(\int p(Y|\theta)V_\theta(Y)\hat{\theta}(Y)dY\right)^2$$

$$= \left(\partial_\theta \int p(Y|\theta)\hat{\theta}(Y)dY\right)^2 = (\partial_\theta\theta)^2 = 1, \tag{3.36}$$

since we may interchange integration by Y and ∂_θ. ■

Eq. (3.31) lends itself to the definition of a one-number summary of the properties of an estimator, namely how closely the inequality is met.

Definition 3.12 (Efficiency) *The statistical efficiency e of an estimator $\hat{\theta}(Y)$ is defined as*

$$e := 1/\det IB. \tag{3.37}$$

The closer e is to 1, the lower the variance of the corresponding estimator $\hat{\theta}(Y)$. For a special class of estimators minimizing loss functions, the following theorem allows us to compute B and e efficiently.

Asymptotic Variance

Theorem 3.13 (Murata, Yoshizawa, Amari [379, Lemma 3]) *Assume that $\hat{\theta}$ is defined by $\hat{\theta}(Y) := \operatorname{argmin}_\theta d(Y, \theta)$ and that d is a twice differentiable function in θ. Then asymptotically, for increasing sample size $m \to \infty$, the variance B is given by $B = Q^{-1}GQ^{-1}$. Here*

$$G_{ij} := \operatorname{cov}_\theta\left[\partial_{\theta_i} d(Y,\theta), \partial_{\theta_j} d(Y,\theta)\right] \text{ and} \tag{3.38}$$

$$Q_{ij} := \mathbf{E}_\theta\left[\partial^2_{\theta_i\theta_j} d(Y,\theta)\right], \tag{3.39}$$

and therefore $e = (\det Q)^2/(\det IG)$.

This means that for the class of estimators defined via d, the evaluation of their asymptotic efficiency can be conveniently achieved via (3.38) and (3.39). For scalar valued estimators $\theta(Y) \in \mathbb{R}$, these expressions can be greatly simplified to

$$I = \int \left(\partial_\theta \ln p(Y|\theta)\right)^2 d\mathrm{P}(Y|\theta), \tag{3.40}$$

$$G = \int \left(\partial_\theta d(Y,\theta)\right)^2 d\mathrm{P}(Y|\theta), \tag{3.41}$$

$$Q = \int \partial^2_\theta d(Y,\theta) d\mathrm{P}(Y|\theta). \tag{3.42}$$

Finally, in the case of continuous densities, Theorem 3.13 may be extended to piecewise twice differentiable continuous functions d, by convolving the latter with a twice differentiable smoothing kernel, and letting the width of the smoothing kernel converge to zero. We will make use of this observation in the next section when studying the efficiency of some estimators.

The current section concludes with the proof that the maximum likelihood estimator meets the Cramér-Rao bound.

Theorem 3.14 (Efficiency of Maximum Likelihood [118, 218, 43]) *The maximum likelihood estimator (cf. (3.18) and (3.19)) given by*

$$\hat{\theta}(Y) := \operatorname*{argmax}_{\theta} \ln p(Y|\theta) = \operatorname*{argmin}_{\theta} \mathcal{L}[\theta] \tag{3.43}$$

is asymptotically efficient ($e = 1$).

To keep things simple we will prove (3.43) only for the class of twice differentiable continuous densities by applying Theorem 3.13. For a more general proof see [118, 218, 43].

Proof By construction, G is equal to the Fisher information matrix, if we choose d according to (3.43). Hence a sufficient condition is that $Q = -I$, which is what we show below. To this end we expand the integrand of (3.42),

$$\partial_\theta^2 d(Y,\theta) = \partial_\theta^2 \ln p(Y|\theta) = \frac{\partial_\theta^2 p(Y|\theta)}{p(Y|\theta)} - \left(\frac{\partial_\theta p(Y|\theta)}{p(Y|\theta)}\right)^2 = \frac{\partial_\theta^2 p(Y|\theta)}{p(Y|\theta)} - V_\theta^2(Y). \tag{3.44}$$

The expectation of the second term in (3.44) equals $-I$. We now show that the expectation of the first term vanishes;

$$\int p(Y|\theta)\frac{\partial_\theta^2 p(Y|\theta)}{p(Y|\theta)} dY = \partial_\theta^2 \int p(Y|\theta) dY = \partial_\theta^2 1 = 0. \tag{3.45}$$

Hence $Q = -I$ and thus $e = Q^2/(IG) = 1$. This proves that the maximum likelihood estimator is asymptotically efficient. ■

It appears as if the best thing we could do is to use the maximum likelihood (ML) estimator. Unfortunately, reality is not quite as simple as that. First, the above statement holds only asymptotically. This leads to the (justified) suspicion that for finite sample sizes we may be able to do better than ML estimation. Second, practical considerations such as the additional goal of sparse decomposition may lead to the choice of a non-optimal loss function.

Finally, we may not know the true density model, which is required for the definition of the maximum likelihood estimator. We can try to make an educated guess; bad guesses of the class of densities, however, can lead to large errors in the estimation (see, e.g., [251]). This prompted the development of robust estimators.

3.4 Robust Estimators

So far, in order to make any practical predictions, we had to *assume* a certain class of distributions from which $\mathrm{P}(Y)$ was chosen. Likewise, in the case of risk functionals, we also assumed that training and test data are identically distributed. This section provides tools to safeguard ourselves against cases where the above