Noname manuscript No.
(will be inserted by the editor)

Regression Analysis: Likelihood, Error and Entropy

Bogdan Grechuk · Michael Zabarankin

Received: date / Accepted: date

Abstract In a regression with independent and identically distributed normal residuals, the log-likelihood function yields an empirical form of the $\mathcal{L}^2$-norm, whereas the normal distribution can be obtained as a solution of differential entropy maximization subject to a constraint on the $\mathcal{L}^2$-norm of a random variable. The $\mathcal{L}^1$-norm and the double exponential (Laplace) distribution are related in a similar way. These are examples of an "inter-regenerative" relationship. In fact, $\mathcal{L}^2$-norm and $\mathcal{L}^1$-norm are just particular cases of general error measures introduced by Rockafellar et al. (2006) on a space of random variables. General error measures are not necessarily symmetric with respect to ups and downs of a random variable, which is a desired property in finance applications where gains and losses should be treated differently. This work identifies a set of all error measures, denoted by $\mathscr{E}$, and a set of all probability density functions (PDFs) that form "inter-regenerative" relationships (through log-likelihood and entropy maximization). It also shows that M-estimators, which arise in robust regression but, in general, are not error measures, form "inter-regenerative" relationships with all PDFs. Remarkably, the set of M-estimators, which are error measures, coincides with $\mathscr{E}$. On the other hand, M-estimators are a particular case of L-estimators that also arise in robust regression. A set of L-estimators which are error measures is identified—it contains $\mathscr{E}$ and the so-called trimmed $\mathcal{L}^p$-norms.

Keywords regression · likelihood · entropy · error measure · M-estimator · L-estimator

Bogdan Grechuk
Department of Mathematics, University of Leicester, LE1 7RH, UK; bg83@leicester.ac.uk

Michael Zabarankin
Department of Mathematical Sciences, Stevens Institute of Technology, Hoboken, NJ 07030, USA;
mzabaran@stevens.edu

1 Introduction

There are at least two approaches to regression analysis: likelihood maximization and error minimization of regression residuals. The first assumes a certain class of probability distributions for the regression residuals and is traditionally used in statistics, whereas the second ponders over a suitable choice of an error measure for the regression residuals and is a customary tool in engineering and risk analysis [39]. Both methods were introduced[1] by Gauss in 1809 [10], who observed that if regression residuals were assumed to be independent and identically distributed (i.i.d.) normal random variables (r.v.'s), then maximization of the log-likelihood function of the regression residuals could be reduced to the least squares problem or, equivalently, to minimization of the $\mathcal{L}^2$-norm of the regression error. In fact, the normal distribution was *introduced* in [10] as the only distribution with such a property. This made the least squares (LS) method as well as the assumption of normally distributed residuals a cornerstone of regression analysis for the past two centuries. (In fact, the LS regression is quite sensitive to outliers—a single outlier may have a drastic impact on regression coefficients [40, pp. 3–5], and there is extensive evidence questioning the assumption on normality of noise in real data [5].) The information theory [21] highlighted another relationship between the $\mathcal{L}^2$-norm and the normal distribution: a normal distribution is a solution of differential entropy maximization [43] with a constraint on the $\mathcal{L}^2$-norm of an r.v. Thus, the log-likelihood function of the normal distribution yields an empirical form of the $\mathcal{L}^2$-norm, whereas the normal distribution can be "recovered" from the maximum entropy principle with a constraint on the $\mathcal{L}^2$-norm:

$$\begin{tikzcd}[column sep=large] \boxed{\begin{array}{c}\text{probability distribution}\\ \text{of regression residuals}\end{array}} \arrow[r, shift left, "\text{log-likelihood function}"] & \boxed{\begin{array}{c}\text{error measure of}\\ \text{regression residuals}\end{array}} \arrow[l, shift left, "\text{entropy maximization}"] \end{tikzcd} \qquad (1)$$

We call (1) an *"inter-regenerative" relationship*.

In fact, the $\mathcal{L}^2$-norm and normal distribution are not the only pair with this remarkable relationship. In 1887, Edgeworth [7] argued[2] that LS regression coefficients are so sensitive to outliers because the residuals are squared, so instead, he suggested to minimize the sum of absolute values of the residuals—the method now known as $\mathcal{L}^1$-regression. (Although coefficients in the $\mathcal{L}^1$-regression are not "immune" to outliers, see e.g. [40, pp. 10–11], the impact of a single outlier in the response variable is not as severe as in the LS regression.) Laplace [26] observed that $\mathcal{L}^1$-regression is equivalent to the likelihood maximization with the double exponential (Laplace) distribution. It turns out that this distribution maximizes the differential entropy subject to a constraint on the $\mathcal{L}^1$-norm [29]. Thus, the $\mathcal{L}^1$-norm and the Laplace distribution is yet another example of (1).

In 1964, Huber [19] proposed to minimize $\sum_i \rho(z_i)$ with respect to regression coefficients, where ρ is a non-constant function and z_i are regression residuals. The

[1] The least squares method was used, although without proofs, by Legendre in 1805 [28], see [17].

[2] The idea to minimize the sum of the absolute deviations of error residuals was first proposed by Boscovich in 1757 [4], see [17].

cases $\rho(t) = t^2$ and $\rho(t) = |t|$ correspond to the LS regression and to the $\mathcal{L}^1$-regression, respectively, while the case $\rho(t) = at^2, t \leqslant 0$, $\rho(t) = bt^2, t \geqslant 0$, with $a > 0, b > 0, a \neq b$, leads to the asymmetric least square (ALS) regression, which is also known as the expectile regression [8,16,44]. This method with different ρ is known as the theory of *M-estimators* [19]. Further, Huber [20] suggested to sum up $\rho(z_i)$ with weights corresponding to the order statistic of z_i, for example, the smallest and the largest residuals could be assigned different weights. This idea leads to the theory of *L-estimators* [20] that generalize M-estimators and that include quantile regression [23] and least median of squares (LMS) regression [42] or least trimmed squares (LTS) regression as particular cases. LMS regression coefficients remain unchanged even if half of all data are outliers. M-estimators and L-estimators remain an active research area, see e.g. [1,18,27,30,32].

The use of M-estimators and L-estimators, as well as other robust estimators, may, however, lead to non-convex optimization for regression coefficients—this is a considerable disadvantage, particularly for large-scale high-dimensional problems. Bernholt [3] suggested an algorithm which computes LMS estimator for n data points in dimension d in time proportional to n^d. Mount et al. [33] offered an $O(n^{d+1})$ algorithm for computing an LTS estimator and showed that the existence of any algorithm which (exactly and deterministically) computes it in time $O(n^k)$ for any $k < d$ would contradict the well-known “hardness of affine degeneracy” conjecture. In real-life applications, particularly with large data sets, LTS regression coefficients can be found by the fast-LTS heuristic [41], but in this case, they are not guaranteed to be optimal.

Rockafellar et al. [39] took the second approach to regression analysis. They introduced general measures of error as nonnegative positively homogeneous convex functionals on a space of r.v.’s, which include the $\mathcal{L}^1$-norm and the $\mathcal{L}^2$-norm, but are not necessarily symmetric with respect to the ups and downs of r.v.’s, and then proposed to minimize a general error measure of regression residuals. For a linear regression, this approach yields convex optimization programs for regression coefficients. Zabarankin and Uryasev [45, Proposition 5.1] showed that entropy maximization subject to a constraint on a general error measure $\mathcal{E}$ is equivalent to entropy maximization subject to two constraints: on the deviation measure projected from $\mathcal{E}$ and on the statistic associated with $\mathcal{E}$.[3] Grechuk and Zabarankin [15] analyzed sensitivity of optimal values of positively homogenous convex functionals in various optimization problems, including linear regression, to noise in the data. The theory of general error measures opens up the possibility for identifying other pairs of error measures and probability distributions that are related by (1). Also, connection between the theory of error measures [39] and the theories of M-estimators [19] and L-estimators [20] is believed to be an open issue.

[3] Rockafellar et al. [38,39] proposed a unifying axiomatic framework for general measures of error, deviation and risk—all of them are positively homogenous convex functionals defined on a space of r.v.’s, see also [37,34].

This work shows that *all possible* pairs of error measures and probability density functions (PDFs) that are related by (1) are determined by

$$\mathcal{E}(X) = \left\|X_{a,b}\right\|_p, \quad X \in \mathcal{L}^p(\Theta),\ a > 0,\ \ b > 0,\ \ p \geqslant 1, \tag{2a}$$

$$f(t) = C\exp\left(-\lambda\, t^p_{a,b}\right), \quad t \in \mathbb{R},\ C > 0,\ \lambda > 0,\ \int_{-\infty}^{\infty} f(t)\,dt = 1, \tag{2b}$$

respectively, where X is an r.v., $\|\cdot\|_p$ is the $\mathcal{L}^p$-norm, and $(\cdot)_{a,b}$ is a function defined by

$$t_{a,b} = a\,[t]_+ + b\,[t]_-, \quad [t]_\pm = \max\{0, \pm t\}. \tag{3}$$

For example, for $a = b = 1$, (2a) simplifies to the $\mathcal{L}^p$-norm $\|X\|_p$, whereas for $p = 1$, $a = 1$ and $b = 1/\alpha - 1$ with $\alpha \in (0,1)$, it is the asymmetric mean absolute error, also known as the Koenker-Bassett error measure used in the quantile regression [23]. The sets of all error measures defined by (2a) and of all PDFs given by (2b) is denoted by $\mathscr{E}$ and $\mathscr{P}$, respectively. If $\mathscr{E}$ is replaced by M-estimators, which, in general, are not error measures in the sense of Rockafellar et al. [39] (positively homogeneous convex nonnegative functionals), then (1) is extended from $\mathscr{P}$ to *all* PDFs, and the set of all M-estimators that are error measures coincides with $\mathscr{E}$, see Figure 1. In fact, M-estimators are a particular case of L-estimators, which are consistent with Huber's theory of robust regression. The set of all error measures which are also L-estimators is denoted by $\mathscr{V}$ and contains $\mathscr{E}$ and so-called trimmed $\mathcal{L}^p$-norms. In addition, this work finds all PDFs that maximize the differential entropy subject to a constraint on an arbitrary law-invariant error measure $\mathcal{E}$.

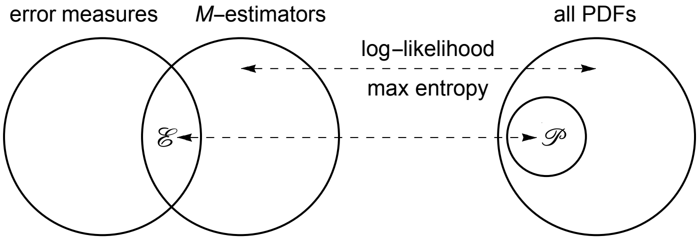

Fig. 1 Relationship between $\mathscr{E}$, $\mathscr{P}$, error measures, and M-estimators.

The rest of the paper is organized into five sections and appendix. Section 2 formulates a general regression problem with error measures and M-estimators and identifies the set of M-estimators which are also error measures. Section 3 discusses entropy maximization subject to constraints on error measures/M-estimators and analyzes correspondence between error measures/M-estimators and maximum entropy distributions. Section 4 extends the results of Sections 2 and 3 for L-estimators. Section 5 concludes the work. Appendix A presents proofs of all the propositions.

2 Log-likelihood function, error measures, and M-estimators

Let $\Theta = (\Omega, \mathcal{M}, \mathbb{P})$ be a probability space, with Ω, $\mathcal{M}$, and $\mathbb{P}$ being a set of elementary events, a σ-algebra over Ω, and a probability measure on $(\Omega, \mathcal{M})$, respectively. A random variable (r.v.) is any measurable function from Ω to $\mathbb{R}$, and $\mathcal{L}^r(\Theta) = \mathcal{L}^r(\Omega, \mathcal{M}, \mathbb{P})$, $r \in [1, \infty]$, is a linear space of r.v.'s with norms $\|X\|_r = (\mathbb{E}[|X|^r])^{1/r}$, $r < \infty$, and $\|X\|_\infty = \operatorname{ess\,sup}|X|$. For an r.v. X, $F_X(x) = \Pr[X \leqslant x]$ and $q_X(s) = \inf\{x | F_X(x) > s\}$ are its cumulative distribution function (CDF) and quantile function, respectively. An r.v. X is *continuous* if $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$ for some function $f_X(t) : \mathbb{R} \to \mathbb{R}^+$, where $\mathbb{R}^+ = [0, +\infty)$, which is called a probability density function (PDF). Θ is *non-trivial* if there exists a non-constant r.v. on Θ, and Θ is *atomless*, if there exists a continuous r.v. on Θ.

Suppose variables $x \in \mathbb{R}^m$ (regressor) and $y \in \mathbb{R}$ (regressant) are related by

$$y = \phi(x;\beta) + z \tag{4}$$

where ϕ is a given function, $\beta \in \mathbb{R}^l$ is an unknown deterministic parameter, and z is a regression error/residual. The regression problem is to find β based on given data $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$.

In statistics, regression residuals $z_i(\beta) = y_i - \phi(x_i;\beta)$, $i = 1, \ldots, n$, are often assumed to be realizations of independent identically distributed (i.i.d.) r.v.'s $Z_i(\beta) \in \mathcal{L}^r(\Theta)$, $i = 1, \ldots, n$, with PDF $f(t) : \mathbb{R} \to \mathbb{R}^+$, so that the likelihood of observing $z_1(\beta), \ldots, z_n(\beta)$ is given by

$$\prod_{i=1}^{n} f(z_i(\beta)). \tag{5}$$

Optimal β is then found by maximizing (5), or equivalently, the logarithm of (5) (log-likelihood function):

$$\max_{\beta \in \mathbb{R}^l} \frac{1}{n} \sum_{i=1}^{n} \ln f(z_i(\beta)), \tag{6}$$

where the multiplier $1/n$ is introduced for convenience.

On the other hand, the objective function in (6) can be considered as the sample analogue of the expected log-likelihood $\mathbb{E}[\ln f(Z;\beta)]$, which is *negative cross entropy*, and the likelihood maximization (6) takes the form

$$\max_{\beta \in \mathbb{R}^l} \mathbb{E}[\ln f(Z(\beta))], \qquad Z(\beta) = Y - \phi(X;\beta). \tag{7}$$

In this case, the functional $\mathcal{E}(Z(\beta)) = -\mathbb{E}[\ln f(Z(\beta))]$ plays the role of a measure for the random error $Z(\beta)$, and the problem (7) can be recast with an arbitrary error measure $\mathcal{E}$:

$$\min_{\beta \in \mathbb{R}^l} \mathcal{E}(Z(\beta)), \qquad Z(\beta) = Y - \phi(X;\beta), \tag{8}$$

which is essentially the approach to regression taken in engineering: find the best fit for the random variable Y in terms of the explanatory random vector $X = (X_1, \ldots, X_m)$.

In general, an error measure is a functional $\mathcal{E} : \mathcal{L}^r(\Theta) \to [0, \infty]$ satisfying the following axioms [39]:

(E1) $\mathcal{E}(0)=0$ but $\mathcal{E}(X)>0$ for nonzero X; also $\mathcal{E}(C)<\infty$ for constant C,
(E2) $\mathcal{E}(\lambda X)=\lambda\mathcal{E}(X)$ for all X and all $\lambda\geqslant 0$ (here $0\,\infty=0$) (*positive homogeneity*),
(E3) $\mathcal{E}(X+Y)\leqslant\mathcal{E}(X)+\mathcal{E}(Y)$ for all X and Y (*subadditivity*),
(E4) $\{X\in\mathcal{L}^r(\Theta)\big|\mathcal{E}(X)\leqslant C\}$ is closed in $\mathcal{L}^r(\Theta)$ for all $C<\infty$ (*lower semicontinuity*).

Loosely speaking, $\mathcal{E}(X)$ is a *nonnegative positively homogeneous convex* functional, which generalizes the notion of norm, but in contrast to a norm is not necessarily symmetric, i.e., in general, $\mathcal{E}(-X)\neq\mathcal{E}(X)$. An error measure $\mathcal{E}$ is called *law invariant* if $\mathcal{E}(X)=\mathcal{E}(Y)$ whenever r.v.'s X and Y have the same distribution.

A broad class of error measures is given by (2a). Comparison of (7) and (8) with (2a) yields (2b)—log-likelihood maximization (6) with (2b) is equivalent to error minimization (8) with (2a).

Example 1 (LS regression) The least squares (LS) regression

$$\min_{\beta\in\mathbb{R}^l}\|Z(\beta)\|_2,\qquad Z(\beta)=Y-\phi(X;\beta),\tag{9}$$

is equivalent to likelihood maximization with a normally distributed regression error.

Example 2 (quantile regression) The quantile regression [23] is equivalent to likelihood maximization with the regression error having the PDF $f(t)=C\exp(-\lambda\,(\alpha\,[t]_+ + (1-\alpha)\,[t]_-))$, $t\in\mathbb{R}$, with $C>0$, $\lambda>0$, and $\alpha\in(0,1)$.

In LS regression (9), a single outlier can substantially alter regression coefficients. Several alternatives have been suggested with better robustness properties. For example, Huber [19] proposed the coefficient vector β in (4) to minimize

$$\min_{\beta\in\mathbb{R}^l}\sum_{i=1}^{n}\rho(z_i(\beta))\tag{10}$$

for some non-constant function $\rho:\mathbb{R}\to\mathbb{R}^+$, where the objective function in (10) is called *M-estimator*. The case $\rho(t)=t^2$ corresponds to the ordinary least square error. Problem (10) is equivalent to (8) with

$$\mathcal{E}(Z)=h(\mathbb{E}[\rho(Z)]),\qquad Z\in\mathcal{L}^r(\Theta),\tag{11}$$

where Z is an r.v. such that $\mathbb{P}[Z=z_i]=1/n$, $i=1,\dots,n$, and $h:\mathbb{R}^+\to\mathbb{R}^+$ is an arbitrary strictly increasing function. For example, with

$$\rho^*(t)=\lambda\,t^p_{a,b},\qquad h(x)=\frac{x^{1/p}}{\lambda},\qquad \lambda>0,\ a>0,\ b>0,\ p\geqslant 1,\tag{12}$$

(11) simplifies to (2a). However, in general, the functional (11) is not an error measure.

Example 3 Let $h(z)=z\in\mathbb{R}^+$ in (11), and $\rho:\mathbb{R}\to\mathbb{R}$ be a convex function, such that $\rho(0)=0$, but $\rho(z)>0$, $z\neq 0$. Then $\mathcal{E}(Z)=\mathbb{E}[\rho(Z)]$ in (11) is a *regular measure of error* [37], i.e., satisfies axioms E1, E3, E4, and

(E5) $\lim\limits_{n\to\infty}\mathcal{E}(Z_n)=0 \implies \lim\limits_{n\to\infty}\mathbb{E}[Z_n]=0$ holds for any sequence $\{Z_n\}_{n=1}^{\infty}$ of r.v.'s.

In general, regular measures of error may not satisfy E2. For example, the asymmetric exponential error $\mathcal{E}(Z) = \mathbb{E}[e^Z - Z - 1]$ satisfies E1 and E3–E5 but not E2, see [37, Example 8].

The following proposition shows that the set of all M-estimators (11), which are error measures, is, in fact, the set $\mathscr{E}$.

Proposition 1 *Let $\mathcal{E} : \mathcal{L}^r(\Theta) \to [0,\infty]$ be an M-estimator (11) defined on non-trivial Θ. Then $\mathcal{E}$ is an error measure if and only if $\mathcal{E} \in \mathscr{E}$.*

Proof See Appendix A.1.

3 Entropy Maximization

Let $\mathcal{C}^1(\Theta) \subset \mathcal{L}^1(\Theta)$ be the set of all r.v.'s, which have finite mean and a PDF, and let $\mathcal{X} \subset \mathcal{C}^1(\Theta)$. Maximization of the differential entropy

$$S(Z) = -\int_{-\infty}^{\infty} f(t) \ln f(t)\, dt$$

can be formulated in a general form:

$$\max_{Z \in \mathcal{X}} S(Z). \tag{13}$$

A set $\mathcal{X}$ is called *law-invariant* if $X \in \mathcal{X}$ implies $Y \in \mathcal{X}$ whenever r.v.'s X and Y have the same distribution.

Proposition 2 *An r.v. $Z^* \in \mathcal{C}^1(\Theta)$ can be a solution to (13) for some convex closed (in $\mathcal{L}^1(\Theta)$) law-invariant set $\mathcal{X}$ if and only if Z^* has a log-concave PDF.*

Proof See Appendix A.2.

Problem (5.4.5) in [45] suggests that maximization of the differential entropy with a constraint on an error measure $\mathcal{E} : \mathcal{L}^r(\Theta) \to [0,\infty]$ of Z:

$$\max_{Z \in \mathcal{L}^r(\Theta)} S(Z) \quad \text{subject to} \quad \mathcal{E}(Z) = 1, \tag{14}$$

can "restore" the PDF of the regression residual. Indeed, if an r.v. Z admits a continuous PDF $f(t) : \mathbb{R} \to \mathbb{R}^+$, then problem (14) with error measure (2a) takes the form

$$\begin{aligned} &\max_{f(t) \geqslant 0} -\int_{-\infty}^{\infty} f(t) \ln f(t)\, dt \\ &\text{subject to} \quad \int_{-\infty}^{\infty} t_{a,b}^p f(t)\, dt = 1, \quad \int_{-\infty}^{\infty} f(t)\, dt = 1, \end{aligned} \tag{15}$$

and Boltzmann's theorem [6, Theorem 11.1.1] yields (2b) with constants $C > 0$ and $\lambda > 0$ to be found from the constraints in (15)—the exact form of f is given by [45, (5.4.8)]

$$f(t) = \frac{1}{(a^{-p} + b^{-p})\, p^{1/p} \Gamma[1 + 1/p]} \exp\left(-\frac{t_{a,b}^p}{p}\right), \qquad t \in \mathbb{R}, \tag{16}$$

where $\Gamma[\cdot]$ is the gamma function. When $a = b = 1$, error measure (2a) takes the form $\mathcal{E}(Z) = \|Z\|_p$ and PDF (16) simplifies to [45, (5.4.9)]

$$f(t) = \frac{1}{2p^{1/p}\Gamma[1+1/p]} \exp\left(-\frac{|t|^p}{p}\right), \qquad t \in \mathbb{R},$$

see Figure 5.2 in [45] for the graph of this PDF for various p.

Thus, given PDF (2b), error measure (2a) follows from the log-likelihood function, and given error measure (2a), PDF (2b) follows from entropy maximization, i.e. (2a) and (2b) form *"inter-regenerative"* relationship (1). This raises the following questions:

(i) *Entropy-error relationship:* For which PDF f does there exist an error measure $\mathcal{E}$ such that f is a maximizer in (14)?
(ii) *Likelihood-error relationship:* For which PDF f does there exist an error measure $\mathcal{E}$ such that (8) yields the same solution as (6)?
(iii) *"Inter-regenerative" relationship:* For which PDF f does there exist an error measure $\mathcal{E}$ which satisfies (i) and (ii) *simultaneously*, i.e., f and $\mathcal{E}$ form (1)?

Questions (i) and (ii) are answered by the following results.

Proposition 3 *A PDF f can be a maximizer in (14) for some law invariant error measure $\mathcal{E}$ if and only if $\log f$ is a concave function.*

Proof See Appendix A.3.

Proposition 4 *Let $\mathcal{E} : \mathcal{L}^r(\Theta) \to [0, \infty]$ be an error measure defined on a non-trivial probability space Θ. If there exists a PDF f such that (6) yields the same solution as (8), then $f \in \mathscr{P}$ and $\mathcal{E} \in \mathscr{E}$.*

Proof See Appendix A.4.

Proposition 4 implies that $\mathscr{P}$ and $\mathscr{E}$ are, in fact, the only sets of PDFs and error measures, respectively, for which the two regression approaches yield the same solution and which form (1).

Example 4 (trimmed $\mathcal{L}^1$-norm) The trimmed $\mathcal{L}^1$-norm (also known as CVaR norm [31]) is the average of the right $(1-\alpha)$-tail of $|Z|$:

$$\mathcal{E}(Z) = \frac{1}{1-\alpha}\int_\alpha^1 q_{|Z|}(s)\,ds = \min_\zeta \left\{\zeta + \frac{1}{1-\alpha}\mathbb{E}[|Z| - \zeta]_+\right\}, \tag{17}$$

where $q_{|Z|}(s)$ is the s-quantile of $|Z|$, is an error measure recently used in regression analysis, see [39]. Since (17) is not in the form (2a), Proposition 4 implies that there is no PDF, for which expected log-likelihood maximization is equivalent to (8) with (17).

Example 5 (mixture of normal distributions) Assume that there are m sources of errors in regression problem (4). Let L be a latent, i.e., unobserved, r.v., such that $L = j$ if and only if the error was caused by source j. Assume that each source produces a normally distributed error, i.e., $z \sim N(\mu_j, \sigma_j)$, if $L = j$, $j = 1, \ldots, m$. Then the (unconditional) density function for z is the mixture of normal distributions

$$f(x) = \sum_{j=1}^{m} \frac{w_j}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right), \tag{18}$$

where $w_j = P[L = j]$, $j = 1, \ldots, m$, which implies that $\sum_{j=1}^{m} w_j = 1$. Let $w_j > 0$, $j = 1, \ldots, m$, i.e., each source of error has a non-zero probability. Parameters $w = (w_1, \ldots, w_m) \in \mathbb{R}^m$, $\mu = (\mu_1, \ldots, \mu_m) \in \mathbb{R}^m$ and $\sigma = (\sigma_1, \ldots, \sigma_m) \in \mathbb{R}^m$ in (18), and $\beta \in \mathbb{R}^l$ in (4) can be found from likelihood maximization through the expected maximization (EM) algorithm [2]. Alternatively, we can minimize some error measure $\mathcal{E}$ of the residuals. However, since f in (18) is not in the form (2b), Proposition 4 implies that there is no error measure for which these two approaches are equivalent. Also, since $\log f$ is not a concave function, Proposition 3 implies that there is no error measure for which f given by (18) is a maximizer in (14).

Next proposition introduces a relationship similar to (1) with M-estimators (11) in place of error measures.

Proposition 5 *Let f^* be an arbitrary PDF. Then* (6) *with the PDF f^* yields the same solution as* (8) *with $\mathcal{E}^*$ in the form* (11) *and $\rho^*(t) = -\ln f^*(t)$. Moreover, f^* can be "restored" from maximization of the differential entropy $S(Z)$ subject to the constraint $\mathcal{E}^*(Z) = c$ for some constant $c \in \mathbb{R}$:*

$$\max_{Z \in \mathcal{L}^r(\Theta)} S(Z) \quad \textit{subject to} \quad \mathcal{E}^*(Z) = c. \tag{19}$$

Proof See Appendix A.5.

4 Generalizations

4.1 L-estimators

Robust alternatives for linear regression use other estimators as well (not just M-estimators). Huber [20] suggested to find regression parameters, $\beta \in \mathbb{R}^l$, in (4) from the optimization problem

$$\min_{\beta \in \mathbb{R}^l} \sum_{i=1}^{n} a_{ni} \rho(z_{(i)}(\beta)), \tag{20}$$

where $\rho : \mathbb{R} \to \mathbb{R}^+$ is a non-constant function, a_{ni} are real coefficients, and $z_{(1)}(\beta), \ldots, z_{(n)}(\beta)$ are the order statistics, i.e., a permutation of $z_1(\beta), \ldots, z_n(\beta)$ such that $z_{(1)}(\beta) \leqslant \ldots \leqslant z_{(n)}(\beta)$. Huber [20] calls $\sum_{i=1}^{n} a_{ni} \rho(z_{(i)}(\beta))$ *L-estimators*. Note that M-estimators are a particular case of L-estimators with $a_{n1} = \cdots = a_{nn} = 1$.

As observed in [20, p. 55], (20) is equivalent to (8) with

$$\mathcal{E}(Z) = \int_0^1 \rho(q_Z(\alpha))M(d\alpha),$$

where M is a signed measure on $(0,1)$, or, equivalently, to (8) with

$$\mathcal{E}(Z) = h\left(\int_0^1 \rho(q_Z(\alpha))M(d\alpha)\right), \tag{21}$$

where $h : \mathbb{R}^+ \to \mathbb{R}^+$ is a strictly increasing function.

An example of L-estimator is a so-called α-trimmed mean ([20, pp. 57–58]) that corresponds to

$$\mathcal{E}(Z) = \frac{1}{1-2\alpha}\int_\alpha^{1-\alpha} \rho(q_Z(\alpha))d\alpha, \quad \alpha \in (0,1/2), \tag{22}$$

where $\rho(z) = z$ or $\rho(z) = |z|$.

There are other versions of robust regression which are similar to (20). They *first* apply a function ρ to residuals and *then* rank them. They correspond to (8) with

$$\mathcal{E}(Z) = h\left(\int_0^1 q_{\rho(Z)}(\alpha)M(d\alpha)\right). \tag{23}$$

A simple example is least median of squares regression

$$\min_{\beta\in\mathbb{R}^l} \operatorname{med}(Z(\beta)^2), \qquad Z(\beta) = Y - \phi(X;\beta),$$

where median $\operatorname{med}(X)$ of an r.v. X is a real number x such that $\Pr[X < x] \leqslant 1/2$ and $\Pr[X > x] \geqslant 1/2$. Coefficients in this regression do not change even if half of the data are outliers, but this regression is much less efficient than (9): more data are required to achieve the same accuracy [42]. Least trimmed squares (LTS) regression has the same robustness level but is more efficient [42, Section 4]. It corresponds to (8) with

$$\mathcal{E}(Z) = \frac{1}{\alpha}\int_0^\alpha q_{|Z|}^2(s)\,ds \tag{24}$$

for some $\alpha \in (0,1)$.

The functionals (22) and (24) are non-convex, and are, therefore, not error measures. The following propositions characterize all error measures which can have either the form (21) or the form (23).

Proposition 6 *Let Θ be an atomless probability space, and let $\mathcal{E} : \mathcal{L}^r(\Theta) \to [0,\infty]$. Then*

(a) $\mathcal{E}$ is an error measure of the form (21) if and only if $\mathcal{E} \in \mathscr{E}$;

(b) $\mathcal{E}$ is an error measure of the form (23) *if and only if it is a particular case of*

$$\mathcal{E}(Z)=\left(\int_0^1 w(\alpha)\,q_{Z_{a,b}}^p(\alpha)\,d\alpha\right)^{1/p}, \quad a>0,\ b>0,\ p\geqslant 1, \tag{25}$$

where $w(\alpha)$ is either a Dirac delta function at 1 or a non-negative non-decreasing function such that $0<\int_0^1 w(\alpha)d\alpha<\infty$.

Proof See Appendix A.6.

Example 6 (trimmed $\mathcal{L}^1$-norm revisited) The trimmed $\mathcal{L}^1$-norm (17) is (25) with $p=1$, $a=b=1$, and $w(s)=(1-s)^{-1}I_{\{s\geqslant\alpha\}}$, where $I_{\{\dots\}}$ is an indicator function equal to 1 if the condition in the curly brackets holds and equal to zero otherwise.

Example 7 (trimmed $\mathcal{L}^p$-norm) Error measure

$$\mathcal{E}(Z)=\left(\frac{1}{1-\alpha}\int_\alpha^1 q_{|Z|}^p(s)\,ds\right)^{1/p}$$

is a hybrid of trimmed $\mathcal{L}^1$-norm (17) and $\mathcal{L}^p$-norm and can be called a *trimmed $\mathcal{L}^p$-norm*. It is (25) with $a=b=1$ and with $w(s)=(1-s)^{-1}I_{\{s\geqslant\alpha\}}$.

A different hybrid of trimmed $\mathcal{L}^1$-norm (17) and $\mathcal{L}^p$-norm is given by

$$\mathcal{E}_{p,\alpha}(Z)=\min_{\zeta}\left\{\zeta+\frac{1}{1-\alpha}\|\,[|Z|-\zeta]_+\|_p\right\}.$$

In fact, $\mathcal{E}_{p,\alpha}(Z)=\mathrm{HMCR}_{p,\alpha}(|Z|)$, where $\mathrm{HMCR}_{p,\alpha}$ is a higher moment coherent risk measure [24]. For $p\in(1,\infty)$ and $\alpha\in(0,1)$, $\mathcal{E}_{p,\alpha}(Z)$ is not of the form (25). Hence, by Proposition 6, it does not belong to family (23) of error measures related to L-estimators.

4.2 Entropy maximization with a constraint on error measure (25)

Example 4 shows that there is no PDF of error residuals such that log-likelihood maximization corresponds to minimization of the trimmed $\mathcal{L}^1$-norm, i.e., the "upper arrow" in (1) does not hold. However, the "lower arrow" still works—differential entropy maximization subject to a constraint on a general error measure was addressed in [45, problem (5.5.4)].

Error measure (25) can be rewritten as

$$\mathcal{E}^p=\int_0^1\left(\int_\alpha^1 q_{Z_{a,b}}^p(s)ds\right)dw(\alpha)+w(0)\int_0^1 q_{Z_{a,b}}^p(s)ds,$$

or

$$\mathcal{E}^p=\int_0^1\left(\frac{1}{1-\alpha}\int_\alpha^1 q_{Z_{a,b}}^p(s)ds\right)(1-\alpha)dw(\alpha)+w(0)\mathbb{E}\left[Z_{a,b}^p\right].$$

The conditional value-at-risk (CVaR) minimization formula [36] (see also (1.4.4) in [45]) yields

$$\frac{1}{1-\alpha}\int_\alpha^1 q^p_{Z_{a,b}}(s)ds = \min_{\zeta(\alpha)}\left(\zeta(\alpha)+\frac{1}{1-\alpha}\mathbb{E}\left[[Z^p_{a,b}-\zeta(\alpha)]_+\right]\right).$$

Then

$$\mathcal{E}^p = \min_{\zeta(\alpha)}\int_0^1\left(\zeta(\alpha)+\frac{1}{1-\alpha}\mathbb{E}\left[[Z^p_{a,b}-\zeta(\alpha)]_+\right]\right)(1-\alpha)dw(\alpha)+w(0)\mathbb{E}\left[Z^p_{a,b}\right],$$

and entropy maximization problem (14) becomes

$$\begin{aligned}&\max_{f(t)\geqslant 0,\,\zeta(\alpha)} -\int_{-\infty}^{\infty} f(t)\ln f(t)\,dt\\ &\text{subject to}\quad \int_{-\infty}^{\infty} f(t)g(t)dt=1,\qquad \int_{-\infty}^{\infty} f(t)\,dt=1,\end{aligned}\tag{26}$$

where

$$g(t)=\int_0^1\left(\zeta(\alpha)+\frac{1}{1-\alpha}[t^p_{a,b}-\zeta(\alpha)]_+\right)(1-\alpha)dw(\alpha)+w(0)t^p_{a,b}.$$

By Boltzmann's theorem [6, Theorem 11.1.1], the maximum-entropy distribution is given by

$$f(t)=c\exp(-\lambda g(t)),$$

where c and λ are positive constants which can be found from the constraints in (26). With $c_2=c\exp\left(-\lambda\int_0^1\zeta(\alpha)(1-\alpha)dw(\alpha)\right)$, we obtain

$$f(t)=c_2\exp\left(-\lambda\left[\int_0^1[t^p_{a,b}-\zeta(\alpha)]_+dw(\alpha)+w(0)t^p_{a,b}\right]\right),\tag{27}$$

where c_2, λ, and $\zeta(\alpha)$ are found from the constraints $\int_0^1 f(t)dt=1$ and $\mathcal{E}(Z)=1$ and the equation

$$\int_{-\infty}^{\infty} I_{\{t^p_{a,b}\leqslant\zeta(\alpha)\}}(t)f(t)dt=\alpha.$$

Example 8 (entropy maximization with trimmed $\mathcal{L}^1$-norm) The entropy maximization problem (14) with the error measure (17) simplifies to

$$\begin{aligned}&\max_{f(t)\geqslant 0,\,\zeta} -\int_{-\infty}^{\infty} f(t)\ln f(t)\,dt\\ &\text{subject to}\quad \int_{-\infty}^{\infty}\left(\zeta+\frac{1}{1-\alpha}[|t|-\zeta]_+\right)f(t)\,dt=1,\qquad \int_{-\infty}^{\infty} f(t)\,dt=1.\end{aligned}$$

As in (27), optimal $f(t)$ has the form

$$f(t)=C\exp(-\lambda[|t|-\zeta]_+),\qquad t\in\mathbb{R},$$

where constants C, λ, and ζ are found from the constraints $\int_0^1 f(t)dt=1$, $\mathcal{E}(Z)=1$, and $\zeta=q_{|Z|}(\alpha)$, so that

$$f(t)=\frac{1}{2}\exp\left(-\frac{1}{1-\alpha}[|t|-\alpha]_+\right),\qquad t\in\mathbb{R}.$$

5 Conclusions

It has long been known that in a regression with independent and identically distributed normal residuals, the log-likelihood function yields an empirical form of the $\mathcal{L}^2$-norm. Conversely, normal distribution can be "restored" as a solution of differential entropy maximization (14) subject to a constraint on the $\mathcal{L}^2$-norm. This is what we call an "inter-regenerative" relationship, i.e., (1). In this work, Proposition 4 shows that an error measure $\mathcal{E}$ can form (1) with some probability density function f if and only if $\mathcal{E}$ belongs to the set $\mathscr{E}$, see (2a). In fact, M-estimators (11), which, in general, are not error measures, form (1) with *all* probability density functions (see Proposition 5). Proposition 1 proves that $\mathscr{E}$ is the only set of error measures in the sense of Rockafellar et al. [39], which are M-estimators, whereas Proposition 6 characterizes the set of all error measures that are L-estimators. In addition, Proposition 2 finds all possible maximum-entropy distributions, for which corresponding entropy maximization problems are convex and law invariant on the space of r.v.'s.

Acknowledgment

We are grateful to the referees for the comments and suggestions, which helped to improve the quality of the paper. The first author thanks the University of Leicester for granting him the academic study leave to do this research.

A Proofs of Propositions 1–6

A.1 Proof of Proposition 1

Since $\mathcal{E}(Z)$ assumes all values in $[0,+\infty)$, the range of h is $[0,+\infty)$, hence it is continuous and $h(0)=0$. This implies that h has a strictly increasing continuous inverse function $h^{-1}:\mathbb{R}^+\to\mathbb{R}^+$, and

$$h^{-1}(\mathcal{E}(Z)) = h^{-1}[h(\mathbb{E}[\rho(Z)])] = \mathbb{E}[\rho(Z)].$$

For constant $Z=t\geqslant 0$,

$$\rho(t) = \mathbb{E}[\rho(t)] = h^{-1}(\mathcal{E}(t)) = h^{-1}(|t|\mathcal{E}(1)).$$

Similarly, $\rho(t)=h^{-1}(|t|\mathcal{E}(-1))$ for $t\leqslant 0$. Consequently, in general,

$$\rho(t) = h^{-1}\left(a\,[t]_+ + b\,[t]_-\right),$$

where $a=\mathcal{E}(1)>0$ and $b=\mathcal{E}(-1)>0$. Thus,

$$\mathcal{E}(Z) = \varphi^{-1}\left(\mathbb{E}\left[\varphi\left(a\,[Z]_+ + b\,[Z]_-\right)\right]\right), \tag{28}$$

where $\varphi = h^{-1}$.

Since $\Theta=(\Omega,\mathcal{M},\mathbb{P})$ is non-trivial, there exists an event $A\in\mathcal{M}$ such that $p=P[A]\in(0,1)$. For any non-negative constants c and d, let Z be an r.v. assuming values $Z(\omega)=c/a\geqslant 0$ and $Z(\omega)=d/a\geqslant 0$ for $\omega\in A$ and $\omega\notin A$, respectively. Then

$$\varphi^{-1}\left[p\varphi(\lambda c)+(1-p)\varphi(\lambda d)\right] = \mathcal{E}(\lambda Z) = \lambda\mathcal{E}(Z) = \lambda\varphi^{-1}\left[p\varphi(c)+(1-p)\varphi(d)\right], \tag{29}$$

for any $\lambda \geqslant 0$. Replacing c and d by $\varphi^{-1}(c)$ and $\varphi^{-1}(d)$, respectively, and applying $\varphi(\cdot)$ to the left-hand and right-hand parts, we obtain

$$p\varphi(\lambda\varphi^{-1}(c)) + (1-p)\varphi(\lambda\varphi^{-1}(d)) = \varphi(\lambda\varphi^{-1}(pc + (1-p)d).$$

Consequently, the function $g(x) = \varphi(\lambda\varphi^{-1}(x))$ satisfies

$$pg(c) + (1-p)g(d) = g(pc + (1-p)d), \quad \forall c, d \geqslant 0. \tag{30}$$

Let

$$\mathcal{A} = \{a \in [0,1] : ag(c) + (1-a)g(d) = g(ac + (1-a)d),\ \forall c, d \geqslant 0\}.$$

By definition, $0 \in \mathcal{A}$ and $1 \in \mathcal{A}$. Also, (30) implies that $pa + (1-p)b \in \mathcal{A}$ whenever $a, b \in \mathcal{A}$, hence $\mathcal{A}$ is a dense subset of $[0,1]$. Finally, $\mathcal{A}$ is closed due to continuity of g, so that $\mathcal{A} = [0,1]$, and g is a linear function. Since $g(0) = \varphi(\lambda\varphi^{-1}(0)) = 0$, there exists a constant $C(\lambda)$ such that

$$\varphi(\lambda\varphi^{-1}(x)) = g(x) = C(\lambda)x, \quad \forall x, \lambda \geqslant 0. \tag{31}$$

Setting $x = \varphi(y)$ in (31), we obtain

$$\varphi(\lambda y) = C(\lambda)\varphi(y), \quad \forall y, \lambda \geqslant 0. \tag{32}$$

Then setting $y = 1$ in (32), we obtain $\varphi(\lambda) = C(\lambda)\varphi(1)$. Consequently, $C(\lambda) = \varphi(\lambda)/\varphi(1)$, and (32) takes the form $\varphi(\lambda y) = \varphi(\lambda)\varphi(y)/\varphi(1)$, $\forall y, \lambda \geqslant 0$. For the function

$$g(x) = \log\frac{\varphi(e^x)}{\varphi(1)},$$

this implies that

$$g(x+y) = \log\frac{\varphi(e^{x+y})}{\varphi(1)} = \log\frac{\varphi(e^x)\varphi(e^y)}{\varphi(1)^2} = g(x) + g(y).$$

Since g is additive, continuous, and $g(0) = 0$, it is linear, i.e., $g(x) = px$ for some constant p. Consequently, $e^{px} = e^{g(x)} = \varphi(e^x)/\varphi(1)$. Finally, with $e^x = y$, we obtain $\varphi(y) = \varphi(1)y^p$, and (28) simplifies to

$$\mathcal{E}(Z) = \left(\mathbb{E}\left[a[Z]_+ + b[Z]_-\right]^p\right)^{1/p}.$$

The condition $p \geqslant 1$ follows from sub-additivity of $\mathcal{E}$.

A.2 Proof of Proposition 2

Proposition 4.7 (b) in [11] implies that if $Z^* \in \mathcal{C}^1(\Theta)$ has a log-concave PDF, then it is a solution to

$$\max_{Z \in \mathcal{C}^1(\Theta)} S(Z) \quad \text{subject to} \quad \mathbb{E}[Z] = \mu, \quad \mathcal{D}(Z) \leqslant 1, \tag{33}$$

for $\mu = \mathbb{E}[Z^*]$ and some law-invariant deviation measure[4] $\mathcal{D}$, and we can set $\mathcal{X} = \{Z \in \mathcal{C}^1(\Theta) \,|\, \mathbb{E}[Z] = \mu,\ \mathcal{D}(Z) \leqslant 1\}$.

Conversely, let $Z^* \in \mathcal{C}^1(\Theta)$ be a solution to (13) for some convex closed law-invariant set $\mathcal{X}$. Then it is a solution to (33) for deviation measure

$$\mathcal{D}(Z) = \sup_{\alpha \in [0,1]} \frac{\mathrm{CVaR}_\alpha^\Delta(Z)}{\mathrm{CVaR}_\alpha^\Delta(Z^*)} \quad \text{for all } Z \in \mathcal{L}^1(\Theta), \tag{34}$$

[4] A deviation measure is a functional $\mathcal{D} : \mathcal{L}^r(\Theta) \to [0, \infty]$ satisfying axioms E2–E4 and such that $\mathcal{D}(Z) = 0$ for constant Z, and $\mathcal{D}(Z) > 0$ otherwise [38]. A deviation measure is called law-invariant if $\mathcal{D}(X) = \mathcal{D}(Y)$ whenever r.v.'s X and Y have the same distribution [12].

where

$$\mathrm{CVaR}_\alpha^\Delta(Z) \equiv \mathbb{E}[Z] - \frac{1}{\alpha}\int_0^\alpha q_Z(s)\,ds, \quad \alpha \in (0,1),$$

$\mathrm{CVaR}_0^\Delta(Z) = \mathbb{E}[Z] - \inf Z$ and $\mathrm{CVaR}_1^\Delta(Z) = \sup Z - \mathbb{E}[Z]$, see [14]. Indeed, if the r.v. Z satisfies the constraints in (33) with $\mathcal{D}$ given by (34), then $\mathbb{E}[Z] = \mu = \mathbb{E}[Z^*]$, and $\mathrm{CVaR}_\alpha^\Delta(Z) \leqslant \mathrm{CVaR}_\alpha^\Delta(Z^*)$ for all $\alpha \in [0,1]$, so that Z dominates Z^* with respect to concave ordering, see Proposition 1 in [14]. Since Z^* has a PDF, the underlying probability space Θ is, by definition, atomless, and part "(a) to (d)" of Corollary 2.61 in [9] along with Lemma 4.2 in [22] implies that $Z \in \mathcal{X}$. Since $Z^* \in \mathcal{C}^1(\Theta)$ is a solution to (13), this yields $S(Z^*) \geqslant S(Z)$, and consequently, Z^* is a solution to (33). Thus, Z^* has a log-concave PDF by Proposition 4.11 in [11].

A.3 Proof of Proposition 3

If $Z^* \in \mathcal{C}^1(\Theta)$ has a log-concave PDF, then it is a solution to (33) for some law-invariant deviation measure $\mathcal{D}$. On the other hand, Proposition 5.1 in [45] shows that problem (33) is equivalent to (14) with an error measure $\mathcal{E}$ such that $\mathcal{D}(Z) = \inf_{C\in\mathbb{R}} \mathcal{E}(Z-C)$, i.e., $\mathcal{D}$ is the deviation measure projected from $\mathcal{E}$. In general, for a given deviation measure $\mathcal{D}$, such an error measure is non-unique and can be determined by

$$\mathcal{E}(Z) = \frac{1}{1+\mu}\left(\mathcal{D}(Z) + |\mathbb{E}[Z]|\right), \tag{35}$$

which is called *inverse projection* of $\mathcal{D}$, see [39]. Thus, Z^* is a solution to (14) with (35).

Conversely, let $Z^* \in \mathcal{C}^1(\Theta)$ be a solution to (14) for some law-invariant error measure $\mathcal{E}$. Then positive homogeneity of $\mathcal{E}$ and relation $S(kZ) = S(Z) + \ln k$, $k > 0$, imply that Z^* is also a maximizer in

$$\max_{Z\in\mathcal{L}^r(\Theta)} S(Z) \quad \text{subject to} \quad \mathcal{E}(Z) \leqslant 1.$$

Since $\{Z\,|\,\mathcal{E}(Z) \leqslant 1\}$ is a convex closed law-invariant set, Z^* has a log-concave PDF by Proposition 2.

A.4 Proof of Proposition 4

If $\mathcal{E}$ and f satisfy the conditions of Proposition 4, then $\mathcal{E}$ and $\rho(t) = -\log(f(t))$ satisfy the conditions of Proposition 1. Consequently, ρ has the form in (12), which implies that $f(t) = e^{-\rho(t)}$, i.e., $f(t)$ is in the form (2b).

A.5 Proof of Proposition 5

Since h is strictly increasing, problem (8) with $\mathcal{E}^*$ is equivalent to minimizing $\mathbb{E}[\rho^*(Z)]$ or to maximizing $\mathbb{E}[\ln(f^*(Z))]$. For an r.v. Z such that $\mathbb{P}[Z = z_i] = 1/n$, $i = 1,\ldots,n$, it reduces to (6).

With $c = h\left(-\int_{-\infty}^{\infty} f^*(t)\ln f^*(t)\,dt\right)$, the constraint $\mathcal{E}^*(Z) = c$ in (19) simplifies to

$$\int_{-\infty}^{\infty} f(t)\ln f^*(t)\,dt = \int_{-\infty}^{\infty} f^*(t)\ln f^*(t)\,dt,$$

which holds for $f = f^*$ and for any $f \neq f^*$ implies that

$$-\int_{-\infty}^{\infty} f(t)\ln f(t)\,dt \leqslant -\int_{-\infty}^{\infty} f(t)\ln f^*(t)\,dt = -\int_{-\infty}^{\infty} f^*(t)\ln f^*(t)\,dt,$$

where the first inequality follows from the non-negativity of relative entropy (Kullback-Leibler divergence between f and f^*), defined as $D_{KL}(f\|f^*) = \int_{-\infty}^{\infty} f(t)\ln\frac{f(t)}{f^*(t)}\,dt \geqslant 0$, see [25].

A.6 Proof of Proposition 6

We first prove the "if" part in (a) and (b). If $\mathcal{E}$ is a particular case of (2a), it is an error measure that can be represented in the form (11), which is (21) with M being a Lebesgue measure on $(0,1)$, and the "if" part in (a) follows. If $\mathcal{E}$ is a particular case of (25), then it can be represented in the form (23) with $M(c,d)=\int_c^d w(\alpha)\,d\alpha$, $0\leqslant c<d\leqslant 1$, $\rho(t)=t_{a,b}^p$, and $h(x)=x^{1/p}$. For $Z\neq 0$, $q_{Z_{a,b}}^p(\alpha)$ is a non-negative non-decreasing function with $\int_0^1 q_{Z_{a,b}}^p(\alpha)\,d\alpha>0$, so that $L=\lim\limits_{\alpha\to 1} q_{Z_{a,b}}^p(\alpha)>0$, and we claim that

$$I=\int_0^1 w(\alpha)\,q_{Z_{a,b}}^p(\alpha)\,d\alpha>0. \tag{36}$$

Indeed, if $w(\alpha)$ is a delta function at 1, (36) reduces to $I=L>0$. Otherwise $\lim\limits_{\alpha\to 1} w(\alpha)>0$, hence $w(\alpha^*)>0$ and $q_{Z_{a,b}}^p(\alpha^*)>0$ for some $\alpha^*<1$, and $I\geqslant\int_{\alpha^*}^1 w(\alpha^*)q_{Z_{a,b}}^p(\alpha^*)=(1-\alpha^*)w(\alpha^*)q_{Z_{a,b}}^p(\alpha^*)>0$.

Inequality $I>0$ implies that $\mathcal{E}(Z)$ is well-defined and satisfies E1. Property E2 is obvious, while E4 is proved for $w(\alpha)=1$ in [38, Proposition 6], and the general case holds by a similar argument. Next, we claim that

$$\mathcal{E}(X+Y)\leqslant\left(\int_0^1 w(\alpha)\,(q_{X_{a,b}}+q_{Y_{a,b}})^p(\alpha)\,d\alpha\right)^{1/p}\leqslant\mathcal{E}(X)+\mathcal{E}(Y) \tag{37}$$

holds for all $X,Y\in\mathcal{L}^r(\Theta)$. Indeed, the second inequality in (37) is a triangle inequality for the $\mathcal{L}^p[0,1]$-norm, and the first one states that

$$\int_0^1 w(\alpha)\,f(\alpha)\,d\alpha\leqslant\int_0^1 w(\alpha)\,g(\alpha)\,d\alpha, \tag{38}$$

for $f(\alpha)=q_{(X+Y)_{a,b}}^p(\alpha)$ and $g(\alpha)=(q_{X_{a,b}}(\alpha)+q_{Y_{a,b}}(\alpha))^p$.

If $f,g\in\mathcal{L}^r[0,1]$ are such that (38) holds for any non-negative non-decreasing $w\in\mathcal{L}^1[0,1]$, we write $g\succcurlyeq f$. The relation $\succcurlyeq$ is

(i) associative;
(ii) monotone, in sense that $f_1(\alpha)\geqslant f_2(\alpha)$ $\forall\alpha\in[0,1]$ implies that $f_1\succcurlyeq f_2$;
(iii) $q_X(\alpha)+q_Y(\alpha)\succcurlyeq q_{X+Y}(\alpha)$ for any r.v.'s $X,Y\in\mathcal{L}^r(\Theta)$ due to sub-additivity of functional $\mathcal{F}(Z)=\int_0^1 w(\alpha)\,q_Z(\alpha)\,d\alpha$, see [13, Proposition 4.3];
(iv) $f_1\succcurlyeq f_2$ is equivalent to $\int_c^1 f_1(\alpha)\,d\alpha\geqslant\int_c^1 f_2(\alpha)\,d\alpha$ for all $c\in(0,1)$, which, in turn, is equivalent to $\int_0^1 u(f_1(\alpha))\,d\alpha\geqslant\int_0^1 u(f_2(\alpha))\,d\alpha$ for all convex increasing u, see [35, Theorem 8]; and
(v) $f_1\succcurlyeq f_2$ implies that $u(f_1)\succcurlyeq u(f_2)$ for any convex increasing function u, which follows from (iv) and the fact that superposition of two convex increasing functions is convex increasing.

Properties (i)–(iii) imply that

$$q_{X_{a,b}}+q_{Y_{a,b}}\succcurlyeq q_{X_{a,b}+Y_{a,b}}\succcurlyeq q_{(X+Y)_{a,b}},$$

and since the function $\xi(z)=z^p$ is convex increasing for $z\geqslant 0$, (38) follows from (v). This finishes the proof of "if" part in (b).

Now we prove the "only if" part. Let $\mathcal{E}$ be an error measure that can be represented in either the form (21) or (23). Since $\mathcal{E}(Z)$ assumes all values in $[0,+\infty)$, h is a strictly increasing continuous function with $h(0)=0$, which has a strictly increasing continuous inverse function $h^{-1}:\mathbb{R}^+\to\mathbb{R}^+$. Applying h^{-1} to both parts of either (21) or (23) and setting $Z=t$, we obtain

$$h^{-1}(\mathcal{E}(t))=\int_0^1 \rho(t)M(d\alpha)=\rho(t)M(0,1),\qquad t\in\mathbb{R}.$$

Consequently, $M(0,1)\neq 0$ and $\rho(t)=\frac{1}{M(0,1)}h^{-1}(\mathcal{E}(t))$. If M and ρ are replaced by $-M$ by $-\rho$, respectively, then $\mathcal{E}$ in (21) remains unchanged. Consequently, without loss of generality, we may assume that $M(0,1)>0$. Positive homogeneity of $\mathcal{E}$ implies that

$$\rho(t)=\frac{1}{M(0,1)}\varphi\left(t_{a,b}\right),$$

where $\varphi = h^{-1}$, $t_{a,b}$ is given by (3), $a = \mathcal{E}(1) > 0$ and $b = \mathcal{E}(-1) > 0$. In particular, both (21) and (23) imply that

$$\mathcal{E}(Z) = \varphi^{-1}\left(\frac{1}{M(0,1)}\int_0^1 q_{\varphi(aZ)}(\alpha)M(d\alpha)\right), \quad Z \geqslant 0, \tag{39}$$

where we used $q_{\varphi(aZ)}(\alpha) = \varphi(q_{aZ}(\alpha))$.

If $M(0,\alpha) = 0$ for all $\alpha < 1$, (21) reduces to $\mathcal{E}(Z) = a[\sup Z]_+ + b[\sup Z]_-$, which is not an error measure (property E1 fails), whereas (23) simplifies to $\mathcal{E}(Z) = \sup(Z_{a,b})$, which is a particular case of (25) with w being the Dirac delta function at 1. Otherwise there exists $\alpha \in (0,1)$ such that $q = M(0,\alpha)/M(0,1) > 0$. Since Θ is atomless, there exists an event $A \in \Theta$ with $P[A] = \alpha$. Let $0 \leqslant c \leqslant d$, and let Z be an r.v. such that $Z(\omega) = c/a$ for $\omega \in A$ and $Z(\omega) = d/a$ for $\omega \notin A$. Then (39) implies that

$$\varphi^{-1}\left[q\varphi(\lambda c) + (1-q)\varphi(\lambda d)\right] = \mathcal{E}(\lambda Z) = \lambda\mathcal{E}(Z) = \lambda\varphi^{-1}\left[q\varphi(c) + (1-q)\varphi(d)\right], \tag{40}$$

for any $\lambda \geqslant 0$. Expression (40) coincides with (29), and the proof of Proposition 1 implies that φ should be in the form $\varphi(y) = \varphi(1)y^p$, $p > 0$. Consequently,

$$h(z) = \left(\frac{z}{\varphi(1)}\right)^{1/p} = h(1)z^{1/p}, \tag{41}$$

and

$$\rho(t) = \frac{\varphi(1)}{M(0,1)}t_{a,b}^p. \tag{42}$$

In particular, (39) simplifies to

$$\mathcal{E}(Z) = \left(\frac{a^p}{M(0,1)}\int_0^1 q_Z(\alpha)^p M(d\alpha)\right)^{1/p}, \quad Z \geqslant 0. \tag{43}$$

Let $0 = \alpha_0 \leqslant \alpha_1 < \alpha_2 < \alpha_3 \leqslant \alpha_4 = 1$ be such that $\alpha_2 - \alpha_1 = \alpha_3 - \alpha_2$, and let

$$M_i = \frac{1}{M(0,1)}\int_{\alpha_{i-1}}^{\alpha_i} M(d\alpha), \qquad i = 1,2,3,4.$$

Since Θ is atomless, there exist events $A, B \in \mathcal{M}$ such that $P[A] = P[B] = \alpha_2$ and $P[A \cap B] = \alpha_1$. Subadditivity of $\mathcal{E}$ implies that

$$\left[\mathcal{E}\left(1 + \epsilon I_{\Omega/A}\right) + \mathcal{E}\left(1 + \epsilon I_{\Omega/B}\right)\right]^p \geqslant \mathcal{E}\left(2 + \epsilon I_{\Omega/A} + \epsilon I_{\Omega/B}\right)^p \quad \forall \epsilon > 0,$$

where I is an indicator function. With (43), this yields

$$2^p\left(M_1 + M_2 + (1+\epsilon)^p(M_3 + M_4)\right) \geqslant 2^p M_1 + (2+\epsilon)^p(M_2 + M_3) + (2+2\epsilon)^p,$$

which simplifies to

$$[(2+2\epsilon)^p - (2+\epsilon)^p]M_3 \geqslant [(2+\epsilon)^p - 2^p]M_2. \tag{44}$$

Dividing both parts by $\epsilon > 0$ and taking limit $\epsilon \to 0^+$, we obtain $p2^{p-1}M_3 \geqslant p2^{p-1}M_2$, or $M_3 \geqslant M_2$. This implies that the measure $M(d\alpha)$ has a non-decreasing density ω on $[0,1]$, which can be the Dirac delta function at the ends of the interval.

By selecting $\alpha_1 = \alpha_2 - \delta$ and $\alpha_3 = \alpha_2 + \delta$ and by taking $\delta \to 0^+$, we can make M_3 arbitrarily close to M_2. Consequently, (44) may hold only if $(2+2\epsilon)^p - (2+\epsilon)^p \geqslant (2+\epsilon)^p - 2^p$. With $\epsilon = 1$, this reduces to $4^p - 2\cdot 3^p + 2^p \geqslant 0$ and implies that $p \geqslant 1$. If $\mathcal{E}$ can be represented in the form (23), this along with (41) and (42) yields (25). Moreover, $\int_0^1 w(\alpha)d\alpha = M[0,1] > 0$. To prove (b), it is left to verify that w is non-negative.

Let $a \geqslant b$ in (25)—the case $a \leqslant b$ is treated similarly. Since Θ is atomless, for every $\alpha \in (0,1/2]$, there exist events $A, B \in \mathcal{M}$ such that $P[A] = P[B] = \alpha$ and $P[A \cap B] = 0$. Subadditivity of $\mathcal{E}$ implies that

$$\mathcal{E}\left(1 - 2I_A\right) + \mathcal{E}\left(1 - 2I_B\right) \geqslant \mathcal{E}\left(2 - 2I_{A\cup B}\right).$$

With (25), this yields

$$2\left(b^p M(0,\alpha)+a^p M(\alpha,1)\right)^{1/p} \geqslant \left((2a)^p M(2\alpha,1)\right)^{1/p},$$

which simplifies to

$$a^p M(\alpha,2\alpha) \geqslant -b^p M(0,\alpha) \quad \forall\alpha\in(0,1/2]. \tag{45}$$

Let $\alpha^* = \sup\{\alpha : w(\alpha) < 0\}$. Since $w(\alpha)$ is non-decreasing, (45) fails for $\alpha = \alpha^*/2$, and consequently, $\alpha^* = 0$. Then $\lim\limits_{\alpha\to 0} M(\alpha,2\alpha) \leqslant \lim\limits_{\alpha\to 0} \alpha w(2\alpha) = 0$, so that $\lim\limits_{\alpha\to 0} M(0,\alpha) \geqslant 0$ by (45), which implies that w has no negative delta function at 0 as well. This finishes the proof of (b).

Finally, suppose that $\mathcal{E}$ is of the form (21). Then an analogue of (43) for negative r.v.'s is given by

$$\mathcal{E}(Z) = \left(\frac{b^p}{M(0,1)}\int_0^1 |q_Z(\alpha)|^p\, M(d\alpha)\right)^{1/p}, \quad Z \leqslant 0. \tag{46}$$

Since $q_{-Z}(\alpha) = -q_Z(1-\alpha)$ for almost all $\alpha\in(0,1)$, (46) can be written as

$$\mathcal{E}(Z') = \left(\frac{b^p}{M(0,1)}\int_0^1 |q_{Z'}(\alpha)|^p\, M'(d\alpha)\right)^{1/p}, \quad Z' \geqslant 0,$$

where $Z' = -Z$ and M' is a measure such that $M'(a,b) = M(1-b,1-a)$ for any interval (a,b). The last expression coincides with (43) and the same argument implies that $M'(d\alpha)$ has a non-decreasing density ω' on $(0,1)$. Since $\omega'(\alpha) = \omega(1-\alpha)$, $\alpha\in(0,1)$, both ω and ω' may be non-decreasing only if ω is constant, which along with (41) and (42) yields (2a) and proves (a).

References

1. Alfons, A., Croux, C., Gelper, S.: Sparse least trimmed squares regression for analyzing high-dimensional large data sets. The Annals of Applied Statistics **7**(1), 226–248 (2013)
2. Bartolucci, F., Scaccia, L.: The use of mixtures for dealing with non-normal regression errors. Computational Statistics & Data Analysis **48**(4), 821–834 (2005)
3. Bernholt, T.: Computing the least median of squares estimator in time o(n^d). In: International Conference on Computational Science and Its Applications, pp. 697–706. Springer (2005)
4. Boscovich, R.J.: De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa. Bononiensi Scientarum et Artum Instituto Atque Academia Commentarii **4**, 353–396 (1757)
5. Box, G.: Non-normality and tests on variances. Biometrika **40**, 318–335 (1953)
6. Cover, T., Thomas, J.: Elements of information theory. John Wiley & Sons (2012)
7. Edgeworth, F.: On observations relating to several quantities. Hermathena **6**(13), 279–285 (1887)
8. Efron, B.: Regression percentiles using asymmetric squared error loss. Statistica Sinica pp. 93–125 (1991)
9. Föllmer, H., Schied, A.: Stochastic Finance, 3 edn. de Gruyter, Berlin New York (2011)
10. Gauss, C.F.: Theoria motus corporum coelestium in sectionibus conicis solem ambientium. sumtibus Frid. Perthes et IH Besser (1809)
11. Grechuk, B., Molyboha, A., Zabarankin, M.: Maximum entropy principle with general deviation measures. Mathematics of Operations Research **34**(2), 445–467 (2009)
12. Grechuk, B., Molyboha, A., Zabarankin, M.: Chebyshev inequalities with law-invariant deviation measures. Probability in the Engineering and Informational Sciences **24**(1), 145–170 (2010)
13. Grechuk, B., Zabarankin, M.: Schur convex functionals: Fatou property and representation. Mathematical Finance **22**(2), 411–418 (2012)
14. Grechuk, B., Zabarankin, M.: Inverse portfolio problem with mean-deviation model. European Journal of Operational Research **234**(2), 481–490 (2014)
15. Grechuk, B., Zabarankin, M.: Sensitivity analysis in applications with deviation, risk, regret, and error measures. SIAM Journal on Optimization (to appear) (2017)
16. Gu, Y., Zou, H.: High-dimensional generalizations of asymmetric least squares regression and their applications. The Annals of Statistics **44**(6), 2661–2694 (2016)

17. Harter, L.: The method of least squares and some alternatives: Part i. International Statistical Review/Revue Internationale de Statistique pp. 147–174 (1974)
18. Hosking, J., Balakrishnan, N.: A uniqueness result for l-estimators, with applications to l-moments. Statistical Methodology **24**, 69–80 (2015)
19. Huber, P.: Robust estimation of a location parameter. The Annals of Mathematical Statistics **35**(1), 73–101 (1964)
20. Huber, P.: Robust statistics. Wiley, New York (1981)
21. Jaynes, E.T.: Information theory and statistical mechanics (notes by the lecturer). In: Statistical Physics 3, vol. 1, p. 181 (1963)
22. Jouini, E., Schachermayer, W., Touzi, N.: Law invariant risk measures have the Fatou property. Advances in mathematical economics **9**, 49–71 (2006)
23. Koenker, R., Bassett Jr, G.: Regression quantiles. Econometrica: journal of the Econometric Society pp. 33–50 (1978)
24. Krokhmal, P.: Higher moment coherent risk measures. Quantitative Finance **7**(4), 373–387 (2007)
25. Kullback, S., Leibler, R.: On information and sufficiency. The Annals of Mathematical Statistics **22**(1), 79–86 (1951)
26. Laplace, P.S.: Traité de mécanique céleste, vol. 2. J. B. M. Duprat, Paris (1799)
27. Lee, W.M., Hsu, Y.C., Kuan, C.M.: Robust hypothesis tests for m-estimators with possibly non-differentiable estimating functions. The Econometrics Journal **18**(1), 95–116 (2015)
28. Legendre, A.M.: Nouvelles méthodes pour la détermination des orbites des comètes. 1. F. Didot (1805)
29. Lisman, J., Van Zuylen, M.: Note on the generation of most probable frequency distributions. Statistica Neerlandica **26**(1), 19–23 (1972)
30. Loh, P.L.: Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. arXiv preprint arXiv:1501.00312 (2015)
31. Mafusalov, A., Uryasev, S.: Cvar (superquantile) norm: Stochastic case. European Journal of Operational Research **249**(1), 200–208 (2016)
32. Morales-Jimenez, D., Couillet, R., McKay, M.: Large dimensional analysis of robust m-estimators of covariance with outliers. IEEE Transactions on Signal Processing **63**(21), 5784–5797 (2015)
33. Mount, D., Netanyahu, N., Piatko, C., Silverman, R., Wu, A.: On the least trimmed squares estimator. Algorithmica **69**(1), 148–183 (2014)
34. Rockafellar, R., Royset, J.: Measures of residual risk with connections to regression, risk tracking, surrogate models, and ambiguity. SIAM J. Optimization **25**(2), 1179–1208 (2015)
35. Rockafellar, R.T., Royset, J.: Random variables, monotone relations, and convex analysis. Mathematical Programming **148**(1-2), 297–331 (2014)
36. Rockafellar, R.T., Uryasev, S.: Conditional value-at-risk for general loss distributions. Journal of Banking and Finance **26**(7), 1443–1471 (2002)
37. Rockafellar, R.T., Uryasev, S.: The fundamental risk quadrangle in risk management, optimization and statistical estimation. Surveys in Operations Research and Management Science **18**(1), 33–53 (2013)
38. Rockafellar, R.T., Uryasev, S., Zabarankin, M.: Generalized deviations in risk analysis. Finance and Stochastics **10**(1), 51–74 (2006)
39. Rockafellar, R.T., Uryasev, S., Zabarankin, M.: Risk tuning with generalized linear regression. Mathematics of Operations Research **33**(3), 712–729 (2008)
40. Rousseeuw, P., Leroy, A.: Robust regression and outlier detection, vol. 589. John Wiley & Sons (2005)
41. Rousseeuw, P., Van Driessen, K.: Computing lts regression for large data sets. Data mining and knowledge discovery **12**(1), 29–45 (2006)
42. Rousseeuw, P.G.: Least median of squares regression. Journal of the American Statistical Association **79**, 871–880 (1984)
43. Shannon, C.E.: A mathematical theory of communication. Bell System Tech **27**, 379423, 623656 (1948)
44. Xie, S., Zhou, Y., Wan, A.: A varying-coefficient expectile model for estimating value at risk. Journal of Business & Economic Statistics **32**(4), 576–592 (2014)
45. Zabarankin, M., Uryasev, S.: Statistical Decision Problems: Selected Concepts and Portfolio Safeguard Case Studies. Springer, Berlin (2014)