Unified robust estimation

Zhu Wang*

The University of Tennessee Health Science Center

Summary

Robust estimation is primarily concerned with providing reliable parameter estimates in the presence of outliers. Numerous robust loss functions have been proposed in regression and classification, along with various computing algorithms. In modern penalised generalised linear models (GLMs), however, there is limited research on robust estimation that can provide weights to determine the outlier status of the observations. This article proposes a unified framework based on a large family of loss functions, a composite of concave and convex functions (CC-family). Properties of the CC-family are investigated, and CC-estimation is innovatively conducted via the iteratively reweighted convex optimisation (IRCO), which is a generalisation of the iteratively reweighted least squares in robust linear regression. For robust GLM, the IRCO becomes the iteratively reweighted GLM. The unified framework contains penalised estimation and robust support vector machine (SVM) and is demonstrated with a variety of data applications.

Key words: CC-estimator; MM algorithm; IRCO; robust; SVM; variable selection.

1. Introduction

Outliers are a small proportion of observations that deviate from the majority and can substantially cause bias in standard estimation methods. This problem has been tackled by robust estimation, which has a long history in statistical methodology research and applications (Hampel et al. 1986; Heritier et al. 2009; Maronna et al. 2019). Denote response variables $y_i$, a $(p + 1)$-dimensional predictor $x_i = (x_{i0}, \ldots, x_{ip})^T$ with the first entry 1, $i = 1, \ldots, n$, and a $(p + 1)$-dimensional coefficient vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$. Robust estimation can be achieved by minimising a loss function

$$\arg \min_{\beta} \sum_{i=1}^{n} \Gamma(y_i, x_i^T \beta),$$

where popular choice of $\Gamma$ in linear regression is the Huber, Andrews or Tukey’s biweight loss. The numerical solutions are typically computed through the so-called iteratively reweighted least squares (IRLS):
$$\arg\min_{\beta} \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2,$$

where weights $w_i$ depend on the loss function $\Gamma$ such that smaller weights are assigned to those observations with larger residuals in magnitude. That is, outliers receive smaller weights. The weights should be understood as $w_i \left( y_i, x_i, x_i^T \beta \right)$ in general. The M-estimators, however, can be defined directly using optimisation problem (2) without the need to introduce the minimisation problem (1).

1.1. Robust logistic regression

For binary outcomes $y_i \in \{0, 1\}$, a robust logistic regression can be obtained by three approaches. First, the parameters can be estimated by a weighted maximum likelihood estimation (WMLE) or equivalently, a weighted minimum negative likelihood estimation

$$\arg\min_{\beta} \sum_{i=1}^{n} w_i s \left( x_i^T \beta, y_i \right)$$

$$s \left( x_i^T \beta, y_i \right) = - (y_i \log p_i(\beta) + (1 - y_i) \log(1 - p_i(\beta)))$$

$$p_i(\beta) = \Pr (y_i = 1|x_i, \beta) = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}.$$

The weights $w_i$ include functions of the deviance and functions of predictors (Green 1984; Carroll & Pederson 1993). A modified method is a weighted estimation equation with a bias correction for consistent estimator (Heritier et al. 2009).

Second, Pregibon (1982) proposed a composite loss function approach given by

$$\arg\min_{\beta} \sum_{i=1}^{n} g \left( s \left( x_i^T \beta, y_i \right) \right),$$

where $g$ is a strictly increasing Huber-type function. This estimator was designed to give less weight to observations poorly fitted by the model. Other functions $g$ have been proposed in Bianco & Yohai (1996), although the estimators may not exist in some applications. To address this issue, Croux & Haesbroeck (2003) proposed different $g$ functions along with a somewhat complex algorithm.

Third, with a focus on prediction, estimation can be achieved by optimising a robust logistic loss function. Park & Liu (2011); Wang (2018) have developed computing algorithms for truncated logistic loss functions, which are Fisher-consistent in classification, meaning that the population minimiser of the loss function leads to the Bayes optimal rule of classification (Lin 2004). However, unlike traditional M-estimation, these approaches fail to retain the weights as a useful diagnostic for the outlier status of the observations.

If the analysis prioritises robust prediction, a natural generalisation of robust logistic regression is sought. An ideal estimation approach should fulfil four criteria:

i. The estimator should be obtained from a loss function satisfying Fisher consistency, which is a fundamental issue from the statistical learning perspective.
ii. A shrinkage estimator can be derived by optimising a penalised loss function. Penalised estimation can improve prediction accuracy and simultaneously conduct parameter estimation and variable selection (Tibshirani 1996; Fan & Li 2001).

iii. The estimation should generate weights to indicate the outlier status of the observations.

iv. The estimator should be computable using a reliable computer algorithm, and it would be advantageous if the algorithm can be generalised to other robust estimation problems.

However, previous robust logistic regression methods only satisfy some of the criteria but not all of them.

1.2. Contribution

We present a novel and unified approach to robust logistic estimation that fulfils all the requirements of the ideal approach mentioned earlier. Our method extends to robust generalised linear models (GLMs) and other related problems, offering a versatile solution. Our contributions can be summarised as follows:

First, we introduce a unified family of robust loss functions, which is a composite of concave and convex functions, known as the CC-family. This family encompasses well-known classical robust loss functions in statistics and data science, such as Huber loss, Andrews loss, biweight loss, robust logistic and hinge loss. Moreover, it also includes a novel robust exponential family.

Second, we propose a new estimation framework that optimises the loss functions within the CC-family. The parameters are estimated using the iteratively reweighted convex optimisation (IRCO) technique, which is a generalisation of the iteratively reweighted least squares (IRLS) used in robust linear regression. The estimated weights provide valuable insights into the outlier status of observations. Additionally, we extend the IRCO method to handle penalised estimation.

Overall, our approach unifies various robust estimation techniques and offers a flexible and efficient solution for various statistical problems.

1.3. Related work

The CC-family encompasses various robust loss functions found in the literature. The concave \( g \) functions within the CC-family include Huber, Andrews and biweight-type functions. In the context of robust logistic regression, the CC-family comprises Huber’s type \( g \) function from Pregibon (1982) and a truncated \( g \) function from Bianco & Yohai (1996). Additionally, a rescaled hinge loss (Xu et al. 2017) also belongs to the CC-family. Notably, the IRCO incorporates the IRLS as a special case for robust linear regression. Moreover, for specific members of the CC-family, the IRCO can be slightly modified to conduct least trimmed squares estimation, and the iteratively reweighted support vector machine (SVM) in Xu et al. (2017) represents a special case of the IRCO. It is worth mentioning that the IRCO offers two approaches for computing weights, with one being simpler than the approach used in Xu et al. (2017).
Alternatively, there is another algorithm for the truncated hinge loss, known as the difference-of-convex (DC) algorithm (Wu & Liu 2007). The DC algorithm decomposes the loss function $\Gamma$ into a difference of two convex functions, whereas the IRCO involves a composite of convex and concave functions. However, the DC algorithm does not update observation weights corresponding to the outlier status, and most CC-family members do not have a simple DC formula except for the truncated loss.

The requirement for a concave function $g$ in the CC-family offers several benefits. For instance, while a composite gradient descent approach can be easily developed to solve a more general composite algorithm and provide greater flexibility in solutions. However, this algorithm lacks the weights as a distinctive characteristic of the outlier status of observations. Moreover, a gradient method may not be the best option in certain scenarios, such as when dealing with the robust hinge loss for SVMs with non-linear kernels like the Gaussian kernel. In contrast, the IRCO for the robust hinge loss effectively corresponds to the iteratively reweighted SVM and can be conveniently implemented using existing software.

The remainder of this article is structured as follows. In Section 2, we present the structure and characteristics of the CC-family. Section 3 details the IRCO for the CC-estimators, explores its convergence properties, and establishes its connections with other algorithms. In Section 4, we illustrate the extensive applications of CC-estimators using both simulated and real data. We showcase a variety of CC-estimators in robust estimation tasks, including regression and GLMs with penalised estimation. In Section 5, we conclude the article with further discussions. The Online Supporting Information provides additional applications, such as the robust SVM, and includes technical proofs.

2. Composite loss functions

The literature has extensively explored a variety of robust loss functions, which are documented in Table 1 (Xu et al. 2017; Wang 2018, 2022; Maronna et al. 2019). These functions can be organised as composite functions, forming the basis of the concave–convex (CC) family.

**Definition 1.** (CC-family). The CC-family contains composite functions $\Gamma = g \circ s$ satisfying the following conditions:

- (i) $g$ is a non-decreasing closed concave function whose domain is the range of function $s$
- (ii) $s$ is convex on $\mathbb{R}$.

The $g$ component, which is concave, robustifies the classical non-robust estimator obtained from the convex $s$ component, such as least squares and negative likelihood functions. The concave property of $g$ is necessary for the IRCO algorithm. Table 2 provides a list of concave components derived from Table 1. Some modifications are required to convert the $g$ of Qloss in Table 1 to ecave, ensuring that the latter is concave with a bounded and continuous derivative. The ecave function is related to erf, the Gaussian error function. Similarly, gcave is constructed from the $g$ of Gloss, ensuring its derivative is bounded and continuous. As shown in Figure 1, all functions, except for heave, are bounded.
Table 1. Composite loss functions with $\sigma > 0$ unless otherwise specified.

| Type         | Loss function $g(s(u))$ | $g(z)$ | $s(u)$ |
|--------------|--------------------------|--------|--------|
| Regression   |                          |        |        |
| Huber        | $\frac{u^2}{2}$ if $|u| \leq \sigma$, $|\sigma| |u| - \frac{\sigma^2}{2}$ if $|u| > \sigma$. | $\frac{z}{2\sigma}$ if $z \leq \sigma^2/2$, $\frac{\sigma(2\sigma)z^2}{2} - \frac{\sigma^2}{2}$ if $z > \sigma^2/2$. | $\frac{\sigma^2}{2}$ |
| Andrews      | $\frac{\sigma}{2\sigma^2} \left(1 - \cos \left(\frac{u}{\sigma}\right)\right)$ if $|u| \leq \sigma \pi$, $\frac{\sigma}{2\sigma^2}$ if $|u| > \sigma \pi$. | $\frac{\sigma}{2\sigma^2} \left(1 - \cos \left(\frac{\sigma^2}{\sigma}z\right)\right)$ if $z \leq \sigma^2 \pi^2/2$, $\frac{\sigma^2}{2}$ if $z > \sigma^2 \pi^2/2$. | $\frac{\sigma^2}{2}$ |
| Biweight     | $1 - \left(1 - \frac{u^2}{\sigma^2}\right)^3 I(|u| \leq \sigma)$ | $1 - \left(1 - \frac{u^2}{\sigma^2}\right)^3 I(z \leq \sigma^2/2)$ | $\frac{\sigma^2}{2}$ |
| CloseR       | $1 - \exp \left(-\frac{u^2}{2\sigma^2}\right)$ | $1 - \exp \left(-\frac{u^2}{2\sigma^2}\right)$ | $\frac{\sigma^2}{2}$ |
| Classification |                          |        |        |
| Closs        | $1 - \exp \left(-\frac{(1-u)^2}{2\sigma^2}\right)$ | $1 - \exp \left(-\frac{\sigma^2}{2\sigma^2}\right)$ | $\frac{(1-u)^2}{2\sigma^2}$ |
| Rhinge       | $1 - \exp \left(-\frac{\max(0,1-u)}{2\sigma}\right)$ | $1 - \exp \left(-\frac{\sigma^2}{2\sigma^2}\right)$ | max $(0, 1 - u)$ |
| Thinge       | $\min(1 - \sigma, \max(0, 1 - u))$, $\min(1 - \sigma, z)$ | $\min(1 - \sigma, z)$ | max $(0, 1 - u)$ |
| Tlogit       | $\min(1 - \sigma, \log(1 + \exp(-u)))$, $\min(1 - \sigma, z)$ | $\min(1 - \sigma, z)$ | $\log(1 + \exp(-u))$ |
| Exp          | $\min(1 - \sigma, \exp(-u))$, $\sigma \leq 0$ | $\min(1 - \sigma, z)$ | $\exp(-u)$ |
| Dlogit       | $\log(1 + \exp(-u)) - \log(1 + \exp(-u - \sigma))$ | $\log \left(\frac{1}{1+z \exp(-\sigma)}\right)$ | $\exp(-u)$ |
| Gloss        | $\frac{1}{(1+z \exp(-\sigma))}, \sigma \geq 1, \sigma > 0$ | $\left(\frac{\sigma}{1+z}\right)^{\sigma}$ | $\exp(-au)$ |
| Qloss        | $1 - \int_{-\infty}^{\frac{Z}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{t^2}{2}\right) \, dt$ | $1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{Z}{\sigma}} \exp(-t^2) \, dt$ | $\frac{\sigma^2}{2}$ |

Table 2. Concave component with $\sigma > 0$.

| Concave | $g(z), z \geq 0$ | Source |
|---------|-----------------|--------|
| hcase   | $\begin{cases} z \sigma(2\sigma)^{z/2} - \frac{\sigma^2}{2} & \text{if } z \leq \sigma^2/2, \\ \frac{z}{2} & \text{if } z > \sigma^2/2. \end{cases}$ | Huber |
| acave   | $\begin{cases} \sigma^2 \left(1 - \cos \left(\frac{\sigma^2}{\sigma}z\right)\right) & \text{if } z \leq \sigma^2 \pi^2/2, \\ 2\sigma^2 & \text{if } z > \sigma^2 \pi^2/2. \end{cases}$ | Andrews |
| bcase   | $\frac{\sigma^2}{2} \left(1 - \left(1 - \frac{z}{\sigma}\right)^3 I(z \leq \sigma^2/2)\right)$ | Biweight |
| ccase   | $\sigma^2 \left(1 - \exp \left(-\frac{z}{\sigma^2}\right)\right)$ | Closs |
| dcase   | $\frac{1}{1+\exp(-\sigma)} \log \left(\frac{1}{1+z \exp(-\sigma)}\right)$ | Dlogit |
| ecase   | $\begin{cases} \frac{2\exp(-\frac{z}{\sigma})}{\sqrt{\pi \sigma^3}} z & \text{if } z \leq \delta, \\ \text{erf} \left(\frac{\sqrt{z}}{\sigma}\right) - \text{erf} \left(\sqrt{\delta}\right) + \frac{2\exp(-\frac{z}{\sigma})}{\sqrt{\pi \sigma^3}} \delta & \text{if } z > \delta. \end{cases}$ | Qloss |
| gcase   | $\begin{cases} \frac{\sigma^{\sigma-1} \gamma}{\gamma} & \text{if } z \leq \delta, \\ \frac{\gamma}{\gamma} \left(\frac{z}{\gamma}\right)^\gamma - \frac{1}{\gamma} \left(\frac{\delta}{\gamma}\right)^\gamma + \frac{\sigma^\gamma}{(1+\sigma^\gamma)^\gamma} & \text{if } z > \delta. \end{cases}$ | Gloss |
| tcave   | $\min(\sigma, z), \sigma \geq 1$ for classification; $\sigma > 0$ otherwise | Truncation |

© 2024 Statistical Society of Australia.
The concave component, along with the derived composite function, is parameterised by $\sigma$, which controls the robustness of the estimation. A smaller value of $\sigma$ allows for more robust estimation. The role of parameter $\sigma$ has been extensively studied in the literature (Wu & Liu 2007; Maronna et al. 2019). The IRCO algorithm in Section 3 will shed light on the impact of $\sigma$ on the estimation process.

Table 3 presents the convex components, which serve as fundamental building blocks in various data analysis theories and applications. For regression problems, the convex component can be Gaussian or $\epsilon$-intensive, which is a crucial device for SVM regression (Hastie, Tibshirani & Friedman 2009). In classification tasks, convex components can be derived from GaussianC, binomial, or hinge loss functions. The GLMs are obtained from the exponential family.

For convenience, Gaussian and binomial losses are separated from the exponential family. In the exponential family, $s(u)$ represents the negative log-likelihood function for certain functions $a(\cdot), b(\cdot)$ and $c(\cdot)$. It is well known that the cumulant function $b(\cdot)$ is
Table 3. Convex component.

| Convex                  | $s(u)$                                      |
|-------------------------|---------------------------------------------|
| Gaussian                | $\frac{u^2}{2}$                             |
| GaussianC               | $\frac{(1-u)^2}{2}$                        |
| Binomial                | $\log(1 + \exp(-u))$                       |
| Exponential family      | $-\left(\frac{2u-b(a_1)}{\beta(\phi)} + c(y, \phi)\right)$ |
| Hinge                   | $\max(0, 1-u)$                             |
| $\epsilon$-insensitive | $\begin{cases} 0 & \text{if } |u| \leq \epsilon, \\ |u| - \epsilon & \text{if } |u| > \epsilon. \end{cases}$ |

convex in its domain (Wainwright & Jordan 2008, Prop. 3.1). Indeed, $s(u)$ is convex in the exponential family. However, it is important to note that $s(u)$ can be negative in certain cases. To construct a valid composite function $g \circ s$ when the domain of $g$ is non-negative, one can make the substitution $s(u)$ with $s(u) - C(y)$, where $C(y)$ is data-dependent and chosen such that $s(u) - C(y) \geq 0$. This can be achieved since $s(u)$ is minimised when $u$ is equivalent to $y$ via a link function in the exponential family. This modification ensures that the composite function remains valid and satisfies the non-negativity constraint of $g$.

Furthermore, by employing common operations with convex functions, it is possible to obtain new members of the CC-family. The corresponding subdifferentials of these functions can be particularly useful in the IRCO algorithm.

**Theorem 1.** Let $\Gamma_1 = g_1 \circ s$ and $\Gamma_2 = g_2 \circ s$ be members of the CC-family $\Omega$ and $c_1, c_2 \geq 0, g = c_1 g_1 + c_2 g_2$. Then $\Gamma = g \circ s \in \Omega$ holds and

$$
\partial(-g(z)) = c_1 \partial(-g_1(z)) + c_2 \partial(-g_2(z))
$$

for any $z$ from $\text{int (dom } g) = \text{int (dom } g_1) \cap \text{int (dom } g_2)$, where $\text{int (dom } g)$ is the interior of domain of $g$.

**Theorem 2.** Let $\Gamma_i = g_i \circ s, i = 1, \ldots, m,$ be members of the CC-family $\Omega, g = \min_{1 \leq i \leq m} \Gamma_i$. Then $\Gamma = g \circ s \in \Omega$ holds. For any $z \in \text{int (dom } g) = \cap_{i=1}^m \text{ int (dom } g_i)$, we have

$$
\partial(-g(z)) = \text{Conv}\{\partial(-g_i(z)) | i \in I(z)\},
$$

where

$$
\text{Conv}\{x_1, \ldots, x_m\} = \left\{ x = \sum_{i=1}^m a_i x_i | a_i \geq 0, \sum_{i=1}^m a_i = 1 \right\},
$$

$I(z) = \{i : g_i(z) = g(z)\}$.

The following properties characterise the robustness of loss functions and are also closely related to the IRCO algorithm.

**Theorem 3.** Assume that $g : \text{range of } s \to \mathbb{R}$, where range of $s$ is open, $g$ and $s$ are twice differentiable, $s'(u) \neq 0$. Then $g$ is concave if and only if for every $u \in \text{dom } s$, the following holds:
\[
\frac{s''(u)}{s'(u)} \Gamma'(u) \geq \Gamma''(u). \tag{7}
\]

For convex function \(s\), since \(s''(u) \geq 0\), (7) is equivalent to
\[
\frac{\Gamma'(u)}{s'(u)} \geq \frac{\Gamma''(u)}{s''(u)},
\]
provided that \(s''(u) \neq 0\). For instance, with \(s(u) = u^2 / 2\), we have for every \(u\),
\[
\frac{\Gamma'(u)}{u} \geq \Gamma''(u).
\]

Note that \(\Gamma'(u)\) is the weight used for M-estimator in robust estimation (Maronna et al. 2019). Likewise, \(g'(s(u)) = \Gamma'(u)/s'(u)\) is the weight in the IRCO.

Theorem 3 is related to the absolute risk aversion for function \(s(u), u \geq 0\):
\[
AR(u) = -\frac{s''(u)}{s'(u)}.
\]

AR is a popular metric in economics for utility function \(s(u)\) that measures preferences over a set of goods and services (Pratt 1964). Assuming non-decreasing function \(s\), we get \(\Gamma'(u) = g'(s(u))s'(u) \geq 0\) for concave function \(g\). Theorem 3 implies that
\[
-\frac{s''(u)}{s'(u)} \leq -\frac{\Gamma''(u)}{\Gamma'(u)}
\]
for \(\Gamma'(u) \neq 0\). Hence, \(\Gamma(u)\) shows globally more risk averse than \(s(u)\) if and only if \(\Gamma(u)\) is a concave transform of \(s(u)\).

Theorem 3 is applicable to many functions in the CC-family, for instance, concave component acave-dcave and gcave (\(\sigma \geq 1\)), and convex component exponential family. The Huber’s type \(g\), however, is only piecewisely twice differentiable. In this case, the following similar results hold.

**Theorem 4.** Assume that \(g\) : range of \(s \to \mathbb{R}\) is continuous, range of \(s = (a, b)\), there is a subdivision \(z_0 = a < z_1 < \cdots < z_k = b\) of \((a, b)\), \(g\) is twice continuously differentiable on each subinterval \((z_{i-1}, z_i), i = 1, \ldots, k\), \(g\) has one-sided derivatives at \(z_1, \ldots, z_k\) satisfying \(D_-g(z_i) \leq D_+g(z_i)\) for \(i = 1, \ldots, k\), \(s\) is twice differentiable, \(s'(u) \neq 0\). Then \(g\) is concave if and only if
\[
\frac{s''(u)}{s'(u)} \Gamma'(u) \geq \Gamma''(u)
\]
holds on each subinterval \((z_{i-1}, z_i), i = 1, \ldots, k\).

Theorem 4 is applicable to the CC-family with concave component hcave, ecave and gcave (\(\sigma < 1\)), and convex component exponential family. With \(s(u) = u^2 / 2, u \geq 0\), \(s\) is non-decreasing. The Gaussian-induced loss functions have larger ARA than that of
Gaussian, provided the ARA exists. For the Huber loss with concave component hcave, simple algebra shows that:

\[ -\frac{s''(u)}{s'(u)} = -\frac{\Gamma''(u)}{\Gamma'(u)}, \quad \text{if } 0 < u < \sigma, \]
\[ -\frac{s''(u)}{s'(u)} < -\frac{\Gamma''(u)}{\Gamma'(u)}, \quad \text{if } u > \sigma. \]

ARA is overlapped with the Gaussian loss when \(0 < u < \sigma\) and greater than the Gaussian when \(u > \sigma\). In other words, we obtain the well-known result: the Huber loss is the same as the Gaussian when \(0 < u < \sigma\) and more robust than the Gaussian otherwise.

Since hinge-type losses do not satisfy a piecewise twice differentiable assumption on the whole domain, Theorems 3 and 4 are not applicable.

2.1. Regression

The CC-family contains Gaussian-induced composite functions, as shown in Figure 2. In addition to classic robust loss functions, new members are introduced from dcave, ecave, and gcave. Figure 2 also includes innovative \(\epsilon\)-insensitive induced loss functions. The composite functions are flatter than their convex counterparts and even become bounded except for hcave, making them more robust to outliers. The derivatives of Gaussian-induced loss functions are shown in Figure 3. With monotone \(\Gamma''\), the M-estimates can break down for high leverage outliers (Maronna et al. 2019, section 5.3). However, except for hcave
Figure 3. Derivatives of Gaussian induced composite loss functions.

(Huber loss), all Gaussian-induced loss functions in Figure 3 are robust to high leverage outliers.

2.2. Classification

For a binary outcome \( y \) taking values \(+1\) and \(-1\), the margin of a classifier \( f \) is denoted by \( u = yf \). Traditional classification problems utilise convex GaussianC, binomial and hinge loss (Hastie, Tibshirani & Friedman 2009). These functions, along with their induced loss functions, are shown in Figure 4. The composite values are normalised such that \( g(s(0)) = 1 \), which effectively requires \( \sigma \geq 1 \) for t\(_{\text{cave}}\). The convex component loss functions are unbounded and cannot control outliers well. On the other hand, the CC-family, except for h\(_{\text{cave}}\) (Huber-type), is bounded, leading to more robust estimation.
The Fisher consistency of margin-based loss functions was initially studied in Lin (2004). In this article, we extend and present additional conditions for Fisher consistency:

1. \( s(u) < s(-u), \; u > 0. \)
2. \( s'(0) < 0. \)
3. \( g : \text{range of } s \rightarrow \mathbb{R} \) is strictly increasing.
4. \( g'(s(0)) \neq 0 \) exists.
5. \( g \circ s \) is a non-increasing function with \( \sigma \geq 1. \)
6. If \( \sigma = 1, \) then \( 1 = g(s(0)) > g(s(1)) \) and \( g(s(0)) = g(s(-1)) \) hold.
7. If \( \sigma > 1, \) then \( g'(s(0)) \neq 0 \) exists.

**Theorem 5.** Assume that \( \Gamma = g \circ s. \) Then for \( Y \in \{-1, 1\}, \Gamma(Yf(X)) \) is Fisher consistent if either of the following two sets of conditions holds:

(i) Conditions 1–4.
(ii) Conditions 2, 5–7.
Conditions 1 and 2 ensure that the function \( s \) is Fisher consistent (Lin 2004). Case (ii) generalises the truncated hinge and logistic loss functions with \( g = \min(\sigma, z) \) (Wu & Liu 2007; Park & Liu 2011). Theorem 5 guarantees that many classification loss functions in the CC-family satisfy the Fisher consistency property. However, one exception is the composite of concave \( t_{\text{cave}} \) and convex GaussianC. This composite function does not satisfy condition 5.

3. Robust estimation

In this section, we present an overview of the estimation problem in the CC-family. We then discuss two different approaches in algorithm design for solving this estimation problem. Next, we provide a detailed description of the IRCO and its convergence results. Finally, we establish connections between the IRCO and the trimmed estimation method.

3.1. Estimation problem

Consider data-dependent convex component \( s(u_i) \) given in Table 3, where

\[
u_i = \begin{cases} y_i - f_i, & \text{for regression,} \\ y_i f_i, & \text{for classification with } y_i \in [-1, 1], \\ f_i, & \text{for exponential family.} \end{cases}
\] (8)

Here \( u_i \) may be seen as \( u_i = u_i(\beta) \) and \( f_i = x_i^T \beta \). Note that \( u_i \) is linked to the linear predictor \( f_i \) via (8), although more complex transformations may be used, such as in the case of non-linear kernels of SVM. A CC-estimator is obtained by finding a solution that minimises the empirical loss \( L(\beta) \) given by

\[
L(\beta) = \frac{1}{n} \sum_{i=1}^{n} \Gamma(u_i(\beta)) = \frac{1}{n} \sum_{i=1}^{n} g(s(u_i(\beta))).
\] (9)

For logistic regression with \( y_i \in \{0, 1\} \), we have

\[
s(u_i) = -y_i x_i^T \beta + \log \left(1 + \exp \left(x_i^T \beta \right)\right),
\]

which is equivalent to the binomial loss in Table 3 with the margin \( u_i = y_i x_i^T \beta, y_i \in [-1, 1] \). Another example is the Poisson regression:

\[
s(u_i) = -y_i x_i^T \beta + \exp (x_i^T \beta).
\]

In many applications, we optimise a penalised loss function \( F : \mathbb{R}^{p+1} \to \mathbb{R} \):

\[
F(\beta) = L(\beta) + \Lambda(\beta),
\] (10)

where

\[
\Lambda(\beta) = \sum_{j=1}^{p} \left( \alpha p_{\lambda}(|\beta_j|) + \lambda \frac{1 - \alpha}{2} \beta_j^2 \right),
\]
0 ≤ α ≤ 1, λ ≥ 0, and \( p_b(|\beta_j|) \) is the penalty function such as the LASSO (Tibshirani 1996) or SCAD (Fan & Li 2001). Minimising the penalised loss function can avoid overfitting, provide shrinkage estimates and conduct variable selection. The loss function (9) is a special case of (10) with \( \Lambda(\beta) = 0 \), i.e., \( \lambda = 0 \).

### 3.2. Algorithm design by the first-order condition of convexity

Suppose \( h \) is a differentiable convex function on its convex domain. Function \( h \), or equivalently, concave function \( g = -h \) has the first-order condition for every \( u, \hat{u} \in \text{dom } g \)

\[
g(u) \leq g(\hat{u}) + g'(\hat{u})(u - \hat{u}).
\]
(11)

Replace \( u \) with \( s(u) \), \( \hat{u} \) with \( s(\hat{u}) \). Thus we have

\[
g(s(u)) \leq g(s(\hat{u})) + g'(s(\hat{u}))(s(u) - s(\hat{u})) = \gamma(u|\hat{u}).
\]
(12)

Then \( \gamma(u|\hat{u}) \) majorises \( \Gamma(u) = g(s(u)) \) at \( \hat{u} \) because we have for every \( u \)

\[
\Gamma(u) \leq \gamma(u|\hat{u}), \quad \Gamma(\hat{u}) = \gamma(\hat{u}|\hat{u}).
\]
(13)

For a non-differentiable function \( g \), similar results hold if the derivative in the first-order condition is replaced with the subgradient. The algorithm follows the majorisation–minimisation (MM) framework (Lange 2016), which is an iterative procedure. Given an estimate \( u^{(k)} \) in the \( k \)th iteration, \( \gamma(u|u^{(k)}) \) is minimised at the \( k + 1 \) iteration to obtain an updated minimiser \( u^{(k+1)} \). This process is repeated until convergence. The MM algorithm generates a descent sequence of estimates:

\[
\Gamma(u^{(k+1)}) \leq \gamma(u^{(k+1)}|u^{(k)}) \leq \gamma(u^{(k)}|u^{(k)}) = \Gamma(u^{(k)}).
\]
(14)

### 3.3. Algorithm design by the Fenchel convex conjugate

Let \( \varphi \) be the convex or Fenchel conjugate of function \( h \) defined by:

\[
\varphi(v) = \sup_{z \in \text{dom } h} (zv - h(z)).
\]

The conjugate \( \varphi \) is convex on \( \text{dom } \varphi \). And conjugate of \( \varphi \) is restored if \( h \) is a closed convex function (Lange 2016, Fenchel–Moreau theorem):

\[
h(z) = \sup_{v \in \text{dom } \varphi} (zv - \varphi(v))
\]

\[
= - \inf_{v \in \text{dom } \varphi} (z(-v) + \varphi(v)).
\]

Let \( h = -g \), where \( g \) is concave. Thus we obtain

\[
g(z) = \inf_{v \in \text{dom } \varphi} (z(-v) + \varphi(v)).
\]

With \( z = s(u) \), we get

\[
g(s(u)) = \inf_{v \in \text{dom } \varphi} (s(u)(-v) + \varphi(v)).
\]
Define
\[ \Gamma(u) = g(s(u)), \quad \zeta(u, v) = s(u)(-v) + \varphi(v). \]  
(15)

Then \( \zeta(u, v) \) majorises \( \Gamma(u) \) at \( \hat{v} \), where \( \hat{v} = \arg\min_v s(u)(-v) + \varphi(v) \). An MM algorithm can be developed to minimise \( \Gamma(u) \) via function \( \zeta(u, v) \) in an alternating scheme. First, given the current value of \( \hat{u} \), we solve \( \hat{v} = \arg\min_v s(\hat{u})(-v) + \varphi(v) \). Second, with the current value of \( \hat{v} \), we minimise \( \zeta(u, \hat{v}) \) with respect to \( u \). This process repeats until convergence. Different from the first-order condition design in Section 3.2, the Fenchel conjugate must be computed. Furthermore, a middle step is required to optimise the \( \zeta(u, \hat{v}) \) in each iteration. However, it will be formally proved in Theorem 6 that the two designs lead to the same solution.

3.4. IRCO

The IRCO to minimise data-driven loss \( F(\beta) \) in (10) is given in Algorithm 1.

**Algorithm 1. IRCO**

1: Initialise \( \beta^{(0)} \) and set \( k = 0 \)
2: repeat
3: Compute \( u_i(\beta^{(k)}) \) in (8) and \( z_i = s(u_i(\beta^{(k)})), i = 1, \ldots, n \)
4: Compute \( v_i^{(k+1)} \) via \( v_i^{(k+1)} \in \partial(-g(z_i)) \) or \( z_i \in \partial \varphi(v_i^{(k+1)}), i = 1, \ldots, n \)
5: Compute \( \beta^{(k+1)} = \arg\min_{\beta} \sum_{i=1}^n s(u_i(\beta))( -v_i^{(k+1)} ) + \Lambda(\beta) \)
6: \( k = k + 1 \)
7: until convergence of \( \beta^{(k)} \)

**Remark 1.** The two approaches to computing the weights in Step 4 correspond to the two algorithm designs in Sections 3.2 and 3.3. Xu et al. (2017) took the approach in Section 3.3 for the composite of the ccave and hinge loss. They derived \( \varphi \) and its derivative to compute the weights. For many applications, the approach in Section 3.2 is much simpler since no middle steps or derivations are required. Furthermore, the weights from the two approaches are the same, thanks to the Fenchel–Moreau theorem. See Theorem 6 and its proof below.

**Remark 2.** Step 4 assumes that \( v_i^{(k+1)} \) exists. This can be justified as follows. If \( v_i^{(k+1)} \) is an interior point of dom \( \varphi \), then \( \partial \varphi( v_i^{(k+1)} ) \) is a non-empty bounded set since conjugate function \( \varphi \) is closed and convex (Nesterov 2004, Theorem 3.1.13). Likewise, if \( -g \) is closed and convex, and \( z_i \) is an interior point of dom \( g \), then \( \partial(-g(z_i)) \) is a non-empty bounded set. Care must be taken on the boundary points. Corresponding to dom \( g = \{ z : z \geq 0 \} \) in Table 2, on boundary point \( z = 0 \), \( g \) must be chosen such that \( \partial(-g(z)) \) is not empty or unbounded. For instance, ecave and gcave \((0 < \sigma < 1)\) are piecewisely constructed to achieve bounded derivative at the origin. For acave, while \( g'(0) \) does not exist, it is simple to choose

\[ g'(0) = \lim_{z \to 0^+} g'(z). \]  
(16)
Table 4. Subdifferential of negative concave component.

| Concave | $\partial(-g(z))$ |
|---------|------------------|
| hcave   | $\begin{cases} -1 & \text{if } z \leq \sigma^2/2, \\ -\sigma(2z)^{-1} & \text{if } z > \sigma^2/2. \end{cases}$ |
| acave   | $\begin{cases} -\frac{\sigma \sin(\sqrt{z}\sigma)}{\sqrt{z}} & \text{if } 0 < z \leq \sigma^2\pi^2/2, \\ -1 & \text{if } z = 0, \\ 0 & \text{if } z > \sigma^2\pi^2/2. \end{cases}$ |
| bcave   | $-\frac{1}{\sigma}\left(2z - \sigma^2\right)^2 \mathbf{1}(z \leq \sigma^2/2)$ |
| cave    | $-\exp\left(-\frac{z}{\sigma^2}\right)$ |
| dcave   | $-\frac{2}{\sigma^2} \exp\left(\frac{z}{\sigma^2}\right)$ |
| ecave   | $\begin{cases} -\frac{1}{\sigma^4}\left(2z - \sigma^2\right)^2 \mathbf{1}(z \leq \sigma^2/2) & \text{if } 0 < z \leq \sigma^2\pi^2/2, \\ -\frac{1}{\sigma^4}\left(2z - \sigma^2\right)^2 \mathbf{1}(z > \sigma^2\pi^2/2). \end{cases}$ |
| tcave   | $\begin{cases} -1 & \text{if } z < \sigma, \\ 0 & \text{if } z > \sigma, \\ [-1, 0] & \text{if } z = \sigma. \end{cases}$

Remark 3. Step 5 amounts to a weighted minimisation problem with weights $-v_i^{(k+1)}$. Since $-g(z)$ is non-increasing convex, we have $v_i^{(k+1)} \leq 0, i = 1, \ldots, n$. Furthermore, $v_i^{(k+1)}$ is a non-decreasing function of $z_i$. See Table 4 and Figure 5. Thereby, ‘clean data’ with small values of $z_i$ will receive larger weights, while outliers with a large value of $z_i$ will receive smaller weights. Note $\sigma$ is suppressed in $g(z)$. For hcave, acave, bcave, ccave and tcave, we obtain $\partial(-g(z, \sigma)) \to -1$ as $\sigma \to \infty$. While a subdifferential is a set by definition, to simplify notations, we interchange between set $\{A\}$ and $A$ when $A$ is the sole element in the set. The relationship between robustness and weights in Table 4 suggests that a larger value $\sigma$ is less robust. Therefore, one may tune the $\sigma$ value from a large value to a small value, that is, from a classical estimator to a robust estimator and select an optimal value of $\sigma$ according to some data-driven criteria. We adopt this procedure in Section 4.

Remark 4. The IRCO is a generalisation of the IRLS to compute M-estimators (Maronna et al. 2019, section 4.5.2). For $\Gamma(u) = g(z), z = s(u) = u^2/2, at the k-th iteration of the IRLS, we compute

$$\arg \min \sum_{i=1}^{n} w_i^{(k+1)}(u_i)u_i^2,$$

where the weights are defined by

$$w_i^{(k+1)}(u_i) = \begin{cases} \Gamma'(u_i)/u_i & \text{if } u_i \neq 0, \\ \Gamma''(0) & \text{if } u_i = 0. \end{cases} \quad (17)$$
Figure 5. Weight function $-\partial(-g(z))$.

It can be shown that $w_i^{(k+1)}(u_i) = -\partial(-g(z_i))$ if $g$ is differentiable at $z_i = s(u_i)$ since we have:

$$-\partial(-g(z_i)) = g'(z_i) = g'(s(u_i)) = \frac{\Gamma'(u_i)}{s'(u_i)} = \frac{\Gamma'(u_i)}{u_i}.$$  

(18)

The remedy in (17) for $u_i = 0$ is the same as (16).

**Remark 5.** Step 5 involves a penalised estimation problem, and we utilise an efficient coordinate descent algorithm, as described in Friedman, Hastie & Tibshirani (2010). In non-convex optimisation, the IRCO typically seeks a local solution, and it is possible to obtain different local solutions with different initial values. Hence, the algorithm may begin with various initial values and determine the best solutions afterwards. For the numerical study in Section 4, we simply initialise $\beta^{(0)} = 0$, and the simulation and data analysis results support this choice.

© 2024 Statistical Society of Australia.
We have obtained convergence results for the IRCO, and the penalty assumptions are provided in the Appendix in the Supporting Information.

**Theorem 6.** Suppose that \( g \) is a concave component in the CC-family, and \( g \) is bounded below.

(i) The loss function values \( F(\beta^{(k)}) \) generated by Algorithm 1 are non-increasing and converge.

(ii) Assume that \( g \) and \( s \) are differentiable, \( \zeta(u, v) = s(u)(-v) + \varphi(v) \) is jointly continuous in \((u, v)\), \( \varphi \) is the conjugate function of \(-g\), \( \nabla L(\beta) = \nabla \ell(\beta | \beta^{(k)}) \), where the surrogate loss is given by

\[
\ell(\beta | \beta^{(k)}) = \sum_{i=1}^{n} \zeta \left( u(\beta), v(\beta^{(k)}) \right),
\]

and \( p_\lambda(\cdot) \) satisfies mild assumptions. Then every limit point of the iterates generated by Algorithm 1 is a Dini stationary point of \( F(\beta) \).

### 3.5. Connection to trimmed estimation

In trimmed least squares (LS), the first step is to compute the residuals from a LS fit. Next, we identify and remove the outliers with large absolute residuals. Finally, we recalculate the LS solution using the remaining observations (Ruppert & Carroll 1980). This estimator can be obtained using the IRCO with a concave \( tcave \) function and the initial estimator being the simple LS solution.

The CC-estimators are also closely related to least trimmed squares (LTS) estimator, which should not be confused with the trimmed LS. Instead of using all \( n \) observations to calculate the regression coefficients, a LTS estimator selects a subset of \( \eta \) observations (where \( \eta < n \)) that result in the smallest sum of squared residuals (least squares) among all possible combinations. See Maronna et al. (2019) and references therein.

To illustrate the connection between Algorithm 1 and LTS, we will explicitly present Algorithm 2 for the concave component \( tcave \) with \( g(z) = \min(\sigma, z) \). This results in the IRCO for the truncation-stationary (IRCOTS) algorithm. In this case, we can obtain the total number of \( z_i \) trimmed by \( \sigma \) in step 4:

\[
\eta^{(k+1)} = \# \left\{ v_i^{(k+1)} = -1, i = 1, \ldots, n \right\}.
\]

The data-driven value of \( \eta^{(k+1)} \) is unspecified but can be computed using the fixed truncation parameter \( \sigma \), which is why it is named truncation-stationary. Next, we modify the IRCOTS algorithm to make the estimator similar to the LTS estimator. Specifically, we adjust Step 4 in Algorithm 2 such that \( \eta^{(k+1)} = \eta \) for all \( k \). This modification allows the location of truncation to change in each iteration.

By doing this, Algorithm 3 seeks a solution for the trimmed estimator as follows:

\[
\hat{\beta} = \arg \min_{\beta} \sum_{i \in H} s(u_i(\beta)) + \Lambda(\beta),
\]

where \( H \subseteq \{1, \ldots, n\} \) and \( |H| = \eta \). This equation represents the trimmed estimator.

Finally, the IRCOTV (IRCO for truncation-varying) algorithm with \( s(u) = u^2/2 \) and LASSO penalty is the same as the algorithm for penalised LTS in Alfons et al. (2013).
Algorithm 2. IRCOTS

1: Initialise $\beta^{(0)}$ and set $k = 0$
2: repeat
3: Compute $u_i(\beta^{(k)})$ in (8) and $z_i = s(u_i(\beta^{(k)})), i = 1, \ldots, n$
4: Compute $v_i^{(k+1)} = -1(z_i \leq \sigma)$
5: Compute $\beta^{(k+1)} = \arg\min_{\beta} \sum_{i=1}^{n} s(u_i(\beta))(v_i^{(k+1)}) + \Lambda(\beta)$
6: $k = k + 1$
7: until convergence of $\beta^{(k)}$

Algorithm 3. IRCOTV

1: Initialise $\beta^{(0)}$ and set $k = 0$
2: repeat
3: Compute $u_i(\beta^{(k)})$ in (8) and $z_i = s(u_i(\beta^{(k)})), i = 1, \ldots, n$
4: Compute $v_i^{(k+1)} = -1(z_i \leq \eta)$, where $z_1 \leq z_2 \cdots \leq z_n$ are ordered statistics, $\eta \leq n$
5: Compute $\beta^{(k+1)} = \arg\min_{\beta} \sum_{i=1}^{n} s(u_i(\beta))(v_i^{(k+1)}) + \Lambda(\beta)$
6: $k = k + 1$
7: until convergence of $\beta^{(k)}$

4. Applications of CC-estimators

We conduct our comparisons using both simulated and real data. The response variables in our experiments include continuous, binary and count data. We choose the robustness parameter $\sigma$ following the guidelines in Remark 3 for Algorithm 1. For penalised estimation, the penalty parameter is determined using data-driven methods described below.

To evaluate the variable selection performance in simulated data, we compute sensitivity (sen) and specificity (spc). Sensitivity measures the proportion of correctly selected predictors among the truly effective predictors, while specificity measures the proportion of correctly non-selected predictors among the truly ineffective predictors. A good estimator should have both sensitivity and specificity close to 1, indicating accurate and precise variable selection.

For more detailed information about the applications and additional results, please refer to the Supporting Information.

4.1. Robust least squares in regression

Example 1. (nonpenalised). Let $y = x^T \beta + \epsilon$, where $\beta = (1.5, 0.5, 1, 1.5, 1)^T$, $\epsilon$ is a $n$-dimensional vector with elements $\epsilon_i$ following a normal distribution with mean 0 and standard deviation 0.5, $i = 1, \ldots, n$. $x_i \sim N_5(\mathbf{0}, \Sigma)$ with $\Sigma_{ij} = 0.5|^{i-j}|$ for $i,j = 1, \ldots, 5$. Training and test data are randomly generated with sample size 100, where training data are used for model estimation, and test data are used to evaluate prediction accuracy. Test data are not contaminated, and contamination mechanisms in the training data follow Alfons et al. (2013):
(1) No contamination
(2) Vertical outliers: 10% of the error terms follow $N(20, 0.5^2)$ instead of $N(0, 0.5^2)$.
(3) Vertical outliers + leverage points: in addition to (2), the 10% contaminated data also have predictor variables distributed as $N(50, 1)$, different from the rest of predictor variables.

Gaussian-induced CC-estimators without penalty are compared with least squares, biweight regression and LTS based on the root mean squared prediction error (RMSE). The average is reported in Table 5 for 100 Monte Carlo simulation runs. The oracle estimator is the true parameter, which provides the best prediction from the simulations. The CC-estimators are comparable with alternative methods for clean data and robust to outliers except for the $hcave$, that is, the Huber estimator. It is well known that the Huber loss is robust to vertical outliers but not leverage points.

Example 2. (penalised). Let $y = x^T \beta + \epsilon$, where $\beta_1 = \beta_7 = 1.5, \beta_2 = 0.5, \beta_4 = \beta_{11} = 1$ and $\beta_j = 0$ otherwise for $j = 1, \ldots, p$. $\epsilon$ is a $n$-dimensional vector with elements $\epsilon_i$ following a normal distribution with mean 0 and standard deviation 0.5, $i = 1, \ldots, n$. $x_i \sim N_p(\mathbf{0}, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$ for $i, j = 1, \ldots, p, p = 50$. We generate random samples and simulation scheme as in Example 1. After training the model with the training data, a separate portion of the data, called the tuning set, is used to fine-tune penalty parameters. The best penalty parameters are chosen to be with the smallest loss values on the tuning set.

Gaussian-induced penalised CC-estimators are computed and are compared with penalised linear regression, robust Huber regression (Yi & Huang 2017) and sparseLTS (Alfons et al. 2013). The results are summarised in Table 6. The penalised CC-estimators are comparable with penalised linear regressions for clean data, and outperform or are comparable with penalised linear regressions, Huber and LTS with outliers. As expected, the Huber loss (hcave) is robust to vertical outliers but not leverage points. The SCAD CC-estimators are better than their corresponding LASSO estimators.

Table 5. RMSE in Example 1.

| Method($\sigma$) | No contamination | Vertical outliers | Vertical+ Leverage |
|------------------|------------------|------------------|-------------------|
| LS               | 0.51             | 2.44             | 3.43              |
| Biweight         | 0.51             | 0.51             | 0.51              |
| LTS              | 0.52             | 0.52             | 0.52              |
| $hcave(1.3)$     | 0.51             | 0.55             | 3.45              |
| $acave(0.9)$     | 0.51             | 0.51             | 0.51              |
| $bave(4.7)$      | 0.51             | 0.51             | 0.51              |
| $cave(1.5)$      | 0.51             | 0.51             | 0.51              |
| $dave(0.5)$      | 0.51             | 0.52             | 0.52              |
| $eave(1.5)$      | 0.52             | 0.52             | 0.52              |
| $gave(1.5)$      | 0.51             | 0.51             | 0.51              |
| $tave(1.0)$      | 0.51             | 0.51             | 0.51              |
| Oracle           | 0.50             | 0.50             | 0.50              |
Table 6. Estimation and prediction in Example 2.

| Method (σ) | No contamination | Vertical outliers | Vertical + Leverage |
|------------|------------------|-------------------|---------------------|
|            | RMSE  | Sen | Spc | RMSE  | Sen | Spc | RMSE  | Sen | Spc |
| LS LASSO   | 0.54  | 1   | 0.76 | 2.96  | 0.63 | 0.84 | 1.73  | 0.98 | 0.50 |
| LS SCAD    | 0.51  | 1   | 0.95 | 2.98  | 0.57 | 0.89 | 1.84  | 0.89 | 0.75 |
| Huber LASSO| 0.54  | 1   | 0.75 | 0.57  | 1.00 | 0.76 | 2.71  | 0.46 | 0.95 |
| SparseLTS  | 0.62  | 1   | 0.92 | 0.58  | 1.00 | 0.90 | 0.58  | 1.00 | 0.89 |
| hcave(0.5)LASSO | 0.54 | 1 | 0.75 | 0.58 | 1.00 | 0.75 | 1.84 | 0.97 | 0.53 |
| hcave(0.5)SCAD | 0.52 | 1 | 0.96 | 0.53 | 1.00 | 0.96 | 1.90 | 0.88 | 0.72 |
| acave(0.9)LASSO | 0.54 | 1 | 0.76 | 0.55 | 1.00 | 0.77 | 0.55 | 1.00 | 0.77 |
| acave(0.9)SCAD | 0.51 | 1 | 0.95 | 0.52 | 1.00 | 0.96 | 0.51 | 1.00 | 0.96 |
| bcave(4.7)LASSO | 0.54 | 1 | 0.76 | 0.55 | 1.00 | 0.77 | 0.55 | 1.00 | 0.77 |
| bcave(4.7)SCAD | 0.51 | 1 | 0.96 | 0.51 | 1.00 | 0.96 | 0.52 | 1.00 | 0.95 |
| ccave(1.5)LASSO | 0.54 | 1 | 0.75 | 0.55 | 1.00 | 0.77 | 0.55 | 1.00 | 0.77 |
| ccave(1.5)SCAD | 0.51 | 1 | 0.95 | 0.51 | 1.00 | 0.96 | 0.51 | 1.00 | 0.96 |
| dcave(0.5)LASSO | 0.54 | 1 | 0.76 | 0.55 | 1.00 | 0.76 | 0.55 | 1.00 | 0.79 |
| dcave(0.5)SCAD | 0.51 | 1 | 0.95 | 0.52 | 1.00 | 0.95 | 0.53 | 1.00 | 0.95 |
| ecave(9.0)LASSO | 0.54 | 1 | 0.74 | 0.55 | 1.00 | 0.76 | 0.54 | 1.00 | 0.82 |
| ecave(9.0)SCAD | 0.52 | 1 | 0.95 | 0.52 | 1.00 | 0.95 | 0.52 | 1.00 | 0.95 |
| gcave(1.5)LASSO | 0.54 | 1 | 0.75 | 0.55 | 1.00 | 0.77 | 0.54 | 1.00 | 0.80 |
| gcave(1.5)SCAD | 0.51 | 1 | 0.96 | 0.51 | 1.00 | 0.96 | 0.54 | 1.00 | 0.95 |
| tcave(2.5)LASSO | 0.54 | 1 | 0.76 | 0.55 | 1.00 | 0.77 | 0.54 | 1.00 | 0.80 |
| tcave(2.5)SCAD | 0.51 | 1 | 0.95 | 0.51 | 1.00 | 0.95 | 0.51 | 1.00 | 0.96 |
| Oracle     | 0.50  | 1   | 1.00 | 0.50  | 1.00 | 1.00 | 0.50  | 1.00 | 1.00 |

4.2. Robust logistic regression

In a survey conducted at a UK hospital, 135 expectant mothers were asked about their decision to breastfeed their babies or not. The survey also collected information on two-level predictive factors (Heritier et al. 2009). We applied binomial-induced CC-estimators, which represent robust logistic regression, to the data and obtained robust weights. Figure 6 displays the robust weights for each individual. Notably, Individuals 3, 11, 14, 53, 63, 75, 90 and 115 received the smallest weights in the robust logistic regression, which confirms the same results as Heritier et al. (2009), but our proposed CC-estimators achieve this using a simpler and more efficient approach.

Interestingly, some individuals showed counterintuitive results when using a logistic regression with large estimated probabilities (greater than or equal to 0.8) for either breastfeeding or not. Despite the high probabilities, these individuals made opposite decisions.

For variable selection, we developed a SCAD logistic regression, which offers sparser estimation than the LASSO estimator when the optimal penalty parameter $\lambda$ is determined using 10-fold cross-validation based on the maximum log-likelihood value. Using the optimal $\lambda$, we computed binomial-induced SCAD CC-estimators and obtained the estimated coefficients for the selected variables, as shown in Table 7.

Comparing the coefficient of $\text{smokenowYes}$ in the penalised logistic regression (which is $-2$), we found that the odds ratio of a desire to breastfeed for a current smoking mother relative to a non-smoking mother is equal to $\text{exp}(-2) = 0.14$. However, the CC-estimators produced coefficients for $\text{smokenowYes}$ that are less than $-2$, indicating that being a smoker during pregnancy has an even larger negative effect according to robust estimation.
Figure 6. Robustness weights of logistic regression for the breastfeeding data.

Table 7. Estimates of robust penalised logistic regression for the breastfeeding data.

| Variable         | logis  | h cave (σ=1.5) | a cave (σ=1.5) | b cave (σ=5) | c cave (σ=2.5) | d cave (σ=3.5) | e cave (σ=7) | g cave (σ=2.5) | t cave (σ=2.2) |
|------------------|--------|----------------|----------------|--------------|----------------|----------------|---------------|----------------|----------------|
| (Intercept)      | 0.10   | -0.20          | 0.32           | 0.33         | 0.35           | 2.71           | 3.27         | -0.70         | -2.27          |
| pregnancyBegin   |        |                |                |              |                |                |               |                |                |
| howfedBreast     |        |                |                |              |                |                |               |                |                |
| howfedfrBreast   | 1.05   | 1.42           | 1.19           | 1.21         | 1.18           | 0.03           | 0.05         | 1.76           | 1.27           |
| partnerPartner   | 0.48   | 0.24           | 0.20           | 0.13         | 0.22           |                |              |                |                |
| smokenowYes      |        |                |                |              |                |                |               |                |                |
| smokebfYes       |        |                |                |              |                |                |               |                |                |
| age              |        |                |                |              |                |                |               |                |                |
| educat           | 0.03   | 0.01           | 0.01           | 0.01         | 0.01           | 0.06           | 0.16         |                |                |
| ethnicNon-white  | 1.94   | 2.49           | 2.52           | 2.64         | 2.48           | 1.16           | 2.45         | 3.25           | 3.59           |

© 2024 Statistical Society of Australia.
Similarly, in all CC-estimators except for dcave, the odds ratios of a desire to breastfeed for a non-White expecting mother relative to a White mother are larger than \( \exp(1.94) = 7 \), which is derived from the penalised logistic regression.

These results highlight the benefits of using robust estimators, such as CC-estimators, in providing more accurate and reliable estimates in the presence of potential outliers and complex relationships in the data.

### 4.3. Robust Poisson regression

In the study of healthcare utilisation among a cohort of 3066 Americans over the age of 50 (Heritier et al. 2009), the outcome of interest was the number of doctor office visits. The survey also contained 24 predictors related to demographic, health needs and economic access. We employed Poisson-induced CC-estimators, also known as robust Poisson regression, to analyse the data. Figure 7 displays the corresponding

![Figure 7. Robustness weights of Poisson regression for the doctor visits data.](image-url)
Table 8. Estimates of robust penalised Poisson regression for the doctor visit data.

| Variable | Poisson | hcase | acave | bcase | ccave | dcave | ecave | gcave | tcave |
|----------|---------|-------|-------|-------|-------|-------|-------|-------|-------|
| (Intercept) | 1.86    | 1.99  | 1.98  | 1.98  | 1.98  | 1.83  | 1.88  | 1.78  | 1.97  |
| age      | $-4 \times 10^{-5}$ | $-4 \times 10^{-5}$ | $-4 \times 10^{-5}$ | $-4 \times 10^{-5}$ | $-4 \times 10^{-5}$ | $-4 \times 10^{-5}$ | $-4 \times 10^{-5}$ | $-4 \times 10^{-5}$ | $-4 \times 10^{-5}$ |
| gender   |         |       |       |       |       |       |       |       |       |
| race     |         |       |       |       |       |       |       |       |       |
| hispan   |         |       |       |       |       |       |       |       |       |
| marital  |         |       |       |       |       |       |       |       |       |
| arthri   | 0.03    | 0.04  | 0.05  | 0.04  | 0.03  | 0.03  | 0.03  | 0.03  | 0.06  |
| cancer   | 0.07    | 0.03  | 0.03  | 0.02  | 0.02  | 0.02  | 0.02  | 0.02  | 0.02  |
| hipress  | 0.12    | 0.11  | 0.08  | 0.12  | 0.13  | 0.05  | 0.07  | 0.07  | 0.08  |
| diabet   | 0.30    | 0.22  | 0.20  | 0.20  | 0.19  | 0.03  | 0.07  | 0.07  | 0.24  |
| lung     | 0.01    | 0.03  | 0.03  | 0.03  | 0.02  | 0.02  | 0.02  | 0.02  | 0.02  |
| heart    | 0.29    | 0.32  | 0.33  | 0.33  | 0.33  | 0.36  | 0.35  | 0.34  | 0.33  |
| stroke   | 0.05    | 0.07  | 0.07  | 0.07  | 0.06  | 0.06  | 0.06  | 0.06  | 0.06  |
| psych    | 0.25    | 0.27  | 0.28  | 0.29  | 0.28  | 0.28  | 0.28  | 0.28  | 0.28  |
| iadla1   | 0.37    | 0.25  | 0.14  | 0.27  | 0.27  | 0.05  | 0.05  | 0.05  | 0.05  |
| iadla2   | 0.68    | 0.44  | 0.37  | 0.39  | 0.40  | 0.36  | 0.36  | 0.36  | 0.36  |
| iadla3   | 0.64    | 0.54  | 0.49  | 0.51  | 0.52  | 0.60  | 0.60  | 0.60  | 0.60  |
| edyears  |         |       |       |       |       |       |       |       |       |
| feduc    |         |       |       |       |       |       |       |       |       |
| meduc    |         |       |       |       |       |       |       |       |       |
| log(income + 1) | 0.04 |       |       |       |       |       |       |       |       |
| insur    | 0.02    |       |       |       |       |       |       |       |       |

robust weights, and interestingly, we observed that the seven smallest weights correspond to subjects with 200, 208, 224, 260, 300, 365 and 750 doctor visits in two years, which aligns with the findings of Heritier et al. (2009) using a more complex M-estimator.

To determine the optimal penalty parameter $\lambda$ for the ordinary SCAD Poisson regression, we conducted a 10-fold cross-validation, maximising the log-likelihood value. Utilising this selected $\lambda$ value, we computed Poisson-induced SCAD CC-estimators. The estimated coefficients of the selected variables are presented in Table 8.

In both the penalised Poisson regression and our Poisson-induced CC-estimators, we observed a negative coefficient for the variable age, suggesting that older patients tend to consume fewer healthcare resources. This finding is consistent with the statistically significant coefficient of $-0.005$ reported by Heritier et al. (2009) using their M-estimator. However, our approach provides a simpler estimation procedure without the need for a complex estimator.

© 2024 Statistical Society of Australia.
5. Discussion

It is important to emphasise that the main objective of this article is to unify various robust loss functions existing in the literature. Additionally, the article aims to extend the application of these loss functions to penalised estimation for shrinkage parameter estimation and variable selection. The article also provides a single computing algorithm that ensures a monotonically decreasing trend in the robust loss values. The IRCO algorithm, which is utilised in this work, holds a practical interpretation for outlier detection. The data-dependent weights employed in the algorithm are linked to outliers, where more extreme observations are assigned smaller weights.

In regression models, when the random error terms have a symmetric distribution, the proposed estimators may hold Fisher consistency with random predictors (Maronna et al. 2019, section 10.11). In the context of GLMs, this class of estimators can be seen as an extension of Pregibon’s work from 1982. However, these estimators do not exhibit Fisher consistency when dealing with random predictors. See Maronna et al. (2019, p. 277) and the cited references for further details on this aspect. Despite its limitations, the proposed approach offers valuable insights into and applications in robust statistical modelling.

This paper proposes a large family of loss functions, the CC-family, which is a composite of concave functions $g(\cdot)$ and convex functions $s(\cdot)$. When applying the CC-family to real applications, the choice of $g(\cdot)$ and $s(\cdot)$ becomes crucial. Selecting appropriate functions can significantly impact model performance. To address this, one may determine an optimal member from the large family of robust loss functions based on model predictive power in applications (Hastie, Tibshirani & Friedman 2009).

In Sections 4.2 and 4.3, we aimed to develop predictive models while identifying potential outliers, comparing the results to those in Heritier et al. (2009). However, it’s important to note that the studies had a limitation: there was no dedicated test dataset to assess and determine optimal models. To overcome this limitation, one could consider splitting the available data into training and test datasets for model evaluation. However, caution should be exercised when comparing the results to Tables 7 and 8 and Figures 6 and 7, as the new models have different sample sizes and possibly different coefficients, model selection results, and outliers.

Although a predictive modelling approach is standard in many cases, we have chosen not to pursue it in this article. Instead, we focus on the development and evaluation of the CC-family and the IRCO algorithm.

We propose potential avenues for further research on CC-estimators. One direction is to explore the efficiency of CC-estimators compared to standard estimators. Specifically, we can investigate the efficiency gains achieved by CC-estimators with concave component and various convex components listed in Tables 2 and 3. Efforts can be made to develop adaptive LASSO CC-estimators, where weighted penalties are prescribed based on the estimated coefficients from a preliminary or initial fit of the model (Zou 2006). The IRCO can be utilised to handle the optimisation problem in adaptive LASSO and examine the properties of the resulting estimators. Oracle properties, similar to those established for adaptive LASSO M-estimators (Smucler & Yohai 2017), could be explored for certain members of the CC-family.

© 2024 Statistical Society of Australia.
Another potential research direction is to consider estimating scale parameters of the exponential family within the CC-family. Robust scale estimators could be developed to address this aspect of the estimation problem (Hampel et al. 1986). These robust scale estimators may prove useful in enhancing the robustness and accuracy of the overall estimation process.

Expanding the convex component of the CC-family opens up possibilities for applying CC-estimators and the IRCO to various statistical applications. For instance, the combination of CC-estimators and decision tree learning-based boosting, a popular toolkit in machine learning (Wang 2021), could lead to novel and effective approaches for handling complex data analysis problems.

In summary, these potential research directions offer exciting opportunities to further explore and extend the CC-family and its associated estimation framework, providing new insights and practical solutions for robust statistical and machine learning applications.

**Supporting information**

Additional supporting information may be found in the online version of this article at http://wileyonlinelibrary.com/journal/anzs.

**Supplementary Information.** Unified robust estimation

**References**

ALFONS, A., CROUX, C., GELPER, S., et al. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7, 226–248.

BIANCO, A.M. & YOHAI, V.J. (1996). Robust estimation in the logistic regression model. In *Robust Statistics, Data Analysis, and Computer Intensive Methods*, pp. 17–34. New York: Springer.

CARROLL, R.J. & PEDERSON, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55, 693–706.

CLARKE, F. (2013). *Functional Analysis, Calculus of Variations and Optimal Control*, Vol. 264. London: Springer Science & Business Media.

CROUX, C. & HAESBROECK, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics & Data Analysis*, 44, 273–295.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.

GREEN, P.J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2), 149–192.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J. & STAHEL, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, Vol. 196. New York: John Wiley & Sons.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J.H. (2009). *The Elements of Statistical Learning*, 2nd edn. New York: Springer-Verlag.

HERITIER, S., CANTONI, E., COPT, S. & VICTORIA-FESER, M.P. (2009). *Robust Methods in Biostatistics*, Vol. 825. Chichester: John Wiley & Sons.

HIRIART-ÚRRUTY, J.B. & LEMARÉCHAL, C. (1993). *Convex Analysis and Minimization Algorithms I: Fundamentals*. Grundlehren der mathematischen Wissenschaften, Vol. 305. New York: Springer-Verlag.

LANGE, K. (2016). *MM Optimization Algorithms*. Philadelphia: SIAM.

LICHTMAN, M. et al. (2013). UCI Machine Learning Repository, 2013. http://archive.ics.uci.edu/ml.
LIN, Y. (2004). A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68, 73–82.

MARONNA, R.A., MARTIN, R.D., YOHAI, V.J. & SALIBIAN-BARRERA, M. (2019). *Robust Statistics: Theory and Methods (with R)*. Hoboken, NJ: John Wiley & Sons.

NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. New York: Springer.

PARK, S.Y. & LIU, Y. (2011). Robust penalized logistic regression with truncated loss functions. *Canadian Journal of Statistics*, 39, 300–323.

PRATT, J. (1964). Risk aversion in the small and in the large. *Econometrica*, 32, 225–243.

PREGIBON, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, 38(2), 485–498.

RUPPERT, D. & CARROLL, R.J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, 75, 828–838.

SMUCLER, E. & YOHAI, V.J. (2017). Robust and sparse estimators for linear regression models. *Computational Statistics & Data Analysis*, 111, 116–130.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

WAINWRIGHT, M.J. & JORDAN, M.I. (2008). Graphical models, exponential families, and variational inference *Foundations and Trends® in Machine Learning*, 1(1-2), 1-305.

WANG, Z. (2018). Robust boosting with truncated loss functions. *Electronic Journal of Statistics*, 12(1), 599–650.

WANG, Z. (2022). MM for Penalized Estimation. *TEST*, 31(1), 54-75.

WANG, Z. (2021). Unified Robust Boosting arXiv:2101.07718 https://arxiv.org/abs/2101.07718. arXiv e-prints

WU, Y. & LIU, Y. (2007). Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102, 974–983.

XU, G., CAO, Z., HU, B.G. & PRINCIPE, J.C. (2017). Robust support vector machines based on the rescaled hinge loss function. *Pattern Recognition*, 63, 139–148.

YI, C. & HUANG, J. (2017). Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics*, 26, 547–557.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.