Neural Networks 50 (2014) 124–141

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Pointwise probability reinforcements for robust statistical inference

Benoît Frénay[*], Michel Verleysen

Machine Learning Group - ICTEAM, Université catholique de Louvain, Place du Levant 3, 1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:
Received 27 March 2013
Received in revised form 14 October 2013
Accepted 14 November 2013

Keywords:
Maximum likelihood
Outliers
Robust inference
Filtering
Cleansing
Probability reinforcements

ABSTRACT

Statistical inference using machine learning techniques may be difficult with small datasets because of abnormally frequent data (AFDs). AFDs are observations that are much more frequent in the training sample that they should be, with respect to their theoretical probability, and include e.g. outliers. Estimates of parameters tend to be biased towards models which support such data. This paper proposes to introduce pointwise probability reinforcements (PPRs): the probability of each observation is reinforced by a PPR and a regularisation allows controlling the amount of reinforcement which compensates for AFDs. The proposed solution is very generic, since it can be used to robustify any statistical inference method which can be formulated as a likelihood maximisation. Experiments show that PPRs can be easily used to tackle regression, classification and projection: models are freed from the influence of outliers. Moreover, outliers can be filtered manually since an abnormality degree is obtained for each observation.

1. Introduction

In statistical inference and machine learning, the goal is often to learn a model from observed data in order to predict a given quantity. In a training sample $\mathbf{x} = (x_1, \ldots, x_n)$, the n observations $x_i \in \mathcal{X}$ are typically assumed to be i.i.d. drawn from the distribution $p(x)$ of the random variable $\mathbf{X}$, whereas the model belongs to a certain parametric family with parameters $\boldsymbol{\theta} \in \Theta$. In particular, many machine learning techniques can be cast as maximum likelihood methods. In this probabilistic framework, learning of the model parameters can be achieved by maximising the data log-likelihood

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{\theta}) \tag{1}$$

where $p(x_i|\boldsymbol{\theta})$ is the probability of the observation x_i under parameters $\boldsymbol{\theta}$. In order to penalise too complex models which could overfit training data, regularisation methods or Bayesian priors can also be used as a complement.

A common problem when the training sample size n is small is that some data may be much more frequent in the training sample that they should be, with respect to their theoretical probability of occurrence $p(x_i)$. These *abnormally frequent data* (AFDs) may pose a threat to statistical inference when maximum likelihood or similar methods are used. Indeed, maximising the log-likelihood corresponds to minimising the Kullback–Leibler divergence between the empirical distribution of observed data and the considered parametric distribution (Barber, 2012), in the hope that the empirical distribution is close to the real (unknown) distribution. Since the empirical probability of AFDs is much larger than their real probability, the parameter estimation is affected and biased towards parameter values which support the AFDs. For example, AFDs are well known to hurt Gaussian distribution fitting. In this paper, a method is proposed to deal with AFDs by considering that it is better to fit for instance 95% of the data well than to fit 100% of the data incorrectly. Notice that outliers are a subclass of AFDs. Indeed, outliers are observations which should theoretically never appear in a training sample, with respect to the parametric model being used (which reflect hypotheses being made about the data generating process). This includes e.g. data which are very far from the mean in Gaussian distribution fitting or data with incorrect labels in classification. Outliers are known to noticeably affect statistical inference. This paper addresses AFDs in general; experiments focus on the specific subclass of outliers.

In many applications, regularisation or Bayesian methods are used to deal with data which are not correctly described by the model, by penalising overly complex models and avoiding overfitting. However, these methods are only suited for the control of model complexity, not for the control of AFD effects. These two problems should be dealt with different methods. Hence, many approaches have been proposed to perform outlier detection (Barnett & Lewis, 1994; Beckman & Cook, 1983; Daszykowski, Kaczmarek, Heyden, & Walczak, 2007; Hawkins, 1980; Hodge & Austin, 2004) and anomaly detection (Chandola, Banerjee, & Kumar, 2009). It is well-known that many statistical inference methods are quite sensitive to outliers, like e.g. linear regression

[*] Corresponding author. Tel.: +32 10 47 81 33; fax: +32 10 47 25 98.
E-mail addresses: benoit.frenay@uclouvain.be (B. Frénay), michel.verleysen@uclouvain.be (M. Verleysen).

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.11.012

(Beckman & Cook, 1983; Cook, 1979; Hadi & Simonoff, 1993), logistic regression (Rousseeuw & Christmann, 2003) or principal component analysis (Archambeau, Delannay, & Verleysen, 2006; Daszykowski et al., 2007; Xu & Yuille, 1995). The approach proposed in this paper relies in part on weighted log-likelihood maximisation, which is often used in the literature to reduce the impact of some of the data (Hu & Zidek, 2002). For example, there exist such algorithms for kernel ridge regression (Jiyan, Guan, & Qun, 2011; Liu, Li, Xu, & Shi, 2011; Suykens, De Brabanter, Lukas, & Vandewalle, 2002; Wen, Hao, & Yang, 2010), logistic regression (Rousseeuw & Christmann, 2003) and principal component analysis (Fan, Liu, & Xu, 2011; Huber, 1981). The main problem with these approaches is that the weights are usually obtained through heuristics. Other methods for linear regression include e.g. M-estimators (Huber, 1964), the trimmed likelihood approach (Hadi & Luceo, 1997) and least trimmed squares (Rousseeuw, 1984; Ruppert & Carroll, 1980). One of the main advantages of the method proposed in this paper is that the observation weights are automatically computed.

AFDs have been widely studied in the classification literature, where labelling errors adversely impact the performances of induced classifiers (Zhu & Wu, 2004). For example, the information gain can be used to detect such AFDs (Guyon, Matic, & Vapnik, 1996). Similarly to the proposed approach, it has also been proposed in the classification literature to limit the influence of each observation during inference, in order to prevent the model parameters to be biased by only a few incorrectly labelled instances. However, each method relies on a different way to limit the contribution of observations which is specific to a given model. For example, instances with large dual weights can be identified as mislabelled for support vector machines (Ganapathiraju, Picone, & State, 2000), on-line learning of perceptrons can be robustified by preventing mislabelled instances to trigger updates too frequently (Kowalczyk, Smola, & Williamson, 2001) and boosting algorithms can impose an upper bound on instance weights (Domingo & Watanabe, 2000). It has also been proposed to associate each observation with a misclassification indicator variable which follows a Bernoulli model (Rekaya, Weigel, & Gianola, 2001), what is closer to the contribution of this paper; the indicators can be used to identify mislabelled observations (Hernandez-Lobato, Hernandez-Lobato, & Dupont, 2011; Zhang, Rekaya, & Bertrand, 2006). The approach proposed in this paper has the advantage of being simple to adapt to specific statistical models and not limited to classification problems.

This paper introduces pointwise probability reinforcements (PPRs), which allow the learner to deal with AFDs in a specific way. The probability of each observation is reinforced by a PPR and a regularisation allows one to control the amount of reinforcement which is awarded to compensate for AFDs. The proposed method is very generic, for it can be applied to any statistical inference method which is the solution of a maximum likelihood problem. Moreover, classical regularisation methods can still be used to further control the model complexity. Eventually, abnormality degrees are obtained, which can be e.g. used to manually screen outliers. In the literature, many outlier detection techniques exist; see e.g. Barnett and Lewis (1994), Beckman and Cook (1983), Hawkins (1980) and Hodge and Austin (2004) for a survey. However, the primary goal of the method proposed in this paper is not only to detect the outliers: the aim is rather to make maximum likelihood estimates less sensitive to observations which are abnormally frequent (including outliers) in the training sample, with respect to their theoretical probability. Consequently, common statistical inference methods like linear regression, kernel ridge regression (a.k.a. least squares support vector machines), logistic regression and principal component analysis are shown to be easily robustified using the proposed approach.

This paper is organised as follows. Section 2 introduces PPRs and motivates their formulation. Section 3 proposes a generic algorithm to compute PPRs and to use them during the statistical inference of model parameters. The proposed algorithm is adapted to several supervised and unsupervised problems in Section 4. It is shown that PPRs allow one to efficiently deal with outliers and Section 5 discusses how to choose the amount of reinforcement to use. The resulting methodology is assessed experimentally for kernel ridge regression in Section 6. Eventually, Section 7 concludes the paper.

2. Pointwise probability reinforcements: definition and concepts

As explained in Section 1, the problem with AFDs is that their empirical probability is much larger than their actual probability. As a consequence, the parameters of models inferred from data with AFDs are biased towards values which overestimate the probability of AFDs. For small training samples, this can have an important impact on the resulting model. For example, in linear regression, outliers can significantly bias the slope and the intercept of an estimated model.

In this paper, it is proposed to deal with AFDs by introducing *pointwise probability reinforcements* (PPRs) $r_i \in \Re^+$. The log-likelihood becomes

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{r}) = \sum_{i=1}^{n} \log \left[p(x_i|\boldsymbol{\theta}) + r_i \right] \tag{2}$$

where each observation x_i is given a PPR r_i which acts as a reinforcement to the probability $p(x_i|\boldsymbol{\theta})$, resulting in a *reinforced probability*. The above log-likelihood is called here the *reinforced log-likelihood*. The PPRs should remain small (or even zero), except for AFDs for which they will compensate for the difference between their large empirical probability and their small probability under a model with parameters $\boldsymbol{\theta}$. The spirit of the proposed method is similar to the one of M-estimators (Huber, 1964) and related approaches (Chen & Jain, 1994; Chuang, Su, & Hsiao, 2000; Liano, 1996). In regression, instead of minimising the sum of the squared residuals, the M-estimator approach consists in minimising another function of the residuals which is less sensitive to extreme residuals. Similarly, PPRs allow one to make maximum likelihood less sensitive to extremely small probabilities. However, there exist many different M-estimators and it is not necessarily easy to choose among them. Moreover, their use is limited to regression. On the contrary, PPRs can be used to robustify maximum likelihood methods for e.g. regression, classification or projection, as shown in Section 4. Moreover, Section 3 shows that PPRs can be easily controlled using regularisation, for example by introducing a notion of sparsity.

Eq. (2) can be motivated by considering methods which are used in the literature to deal with outliers. In classification, data consists of pairs $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ where x_i is a vector of observed feature values and y_i is the observed label. Label noise occurs when a few data have incorrect labels (e.g. false positives in medical diagnosis). In such a case, Lawrence and Schölkopf (2001) introduce a labelling error probability π_e which can be used to write

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \pi_e; \mathbf{x}, \mathbf{y}) &= \sum_{i=1}^{n} \log \left[(1-\pi_e)\, p(y_i|x_i, \boldsymbol{\theta}) \right.\\ &\quad \left. + \pi_e \left(1 - p(y_i|x_i, \boldsymbol{\theta})\right) \right]\\ &= \sum_{i=1}^{n} \log \left[p(y_i|x_i, \boldsymbol{\theta}) + \frac{\pi_e}{1-2\pi_e} \right]\\ &\quad + n \log \left[1 - 2\pi_e\right].\end{aligned} \tag{3}$$

Since π_e is small (incorrect labels are not majority), it follows that $\log[1-2\pi_e] \approx 0$ and the log-likelihood (3) can be approximated with (2).

Another possibility to deal with outliers (Aitkin & Wilson, 1980; Eskin, 2000) consists in assuming that data actually come from a mixture of two processes: the actual process of interest and a garbage process generating outliers. The log-likelihood becomes

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\theta}, \pi_g, \boldsymbol{\theta}_g; \mathbf{x}\right) &= \sum_{i=1}^{n} \log\left[\left(1-\pi_g\right) p\left(x_i|\boldsymbol{\theta}\right) + \pi_g p\left(x_i|\boldsymbol{\theta}_g\right)\right] \\ &= \sum_{i=1}^{n} \log\left[p\left(x_i|\boldsymbol{\theta}\right) + \frac{\pi_g}{1-\pi_g} p\left(x_i|\boldsymbol{\theta}_g\right)\right] \\ &\quad + n\log\left[1-\pi_g\right] \qquad (4)\end{aligned}$$

where π_g is the prior probability of garbage patterns and $p\left(x|\boldsymbol{\theta}_g\right)$ is the probability of a garbage pattern x. Since garbage patterns are assumed to be in minority, it follows that π_g is small and that $\log\left[1-\pi_g\right] \approx 0$. Therefore, the above log-likelihood can be also approximated with (2).

It should be stressed that PPRs are not probabilities. Indeed, they can take any positive value, even if they should normally remain small. PPRs are intended to reinforce the probability $p\left(x_i|\boldsymbol{\theta}\right)$ of AFDs whose empirical probability is too large with respect to their true probability $p(x_i)$. This way, the probability $p\left(x_i|\boldsymbol{\theta}\right)$ can remain small for AFDs, what is natural since their true probability is also small. Using PPRs, maximum likelihood parameter estimation is expected to be less sensitive to AFDs and to provide more accurate parameter estimates. The advantage of using PPRs over the two above approaches discussed in Aitkin and Wilson (1980), Eskin (2000) and Lawrence and Schölkopf (2001) is that it is no longer necessary to modify the data generation model. In other words, AFDs do not have to fit into the parametric model which is learnt for prediction. Indeed, a non-parametric method is proposed in Section 3 to compute PPRs using regularisation. Of course, there is no such thing as a free lunch and it is necessary to control the amount of reinforcement which is given. This point is further discussed in detail in Section 5.

3. Statistical inference with pointwise probability reinforcements

This section shows how to perform maximum likelihood statistical inference with PPRs using a generic two-step iterative algorithm. This algorithm is adapted to supervised and unsupervised problems in Section 4.

3.1. Non-parametric pointwise probability reinforcements

Without restriction on the PPRs, the reinforced log-likelihood (2) is unbounded. Indeed, one can simply choose large PPR values and obtain an arbitrary large reinforced log-likelihood, whatever the choice of the model parameters $\boldsymbol{\theta}$. A first solution to this problem is to assume a parametric form $r_i = r\left(x_i|\boldsymbol{\theta}_r\right)$ for the PPRs, where $\boldsymbol{\theta}_r$ are fixed parameters. However, this solution requires prior knowledge on the reinforcement $r\left(x_i|\boldsymbol{\theta}_r\right)$ which is necessary for a given observation x_i. This may depend on the problem which is addressed and the model which is used to solve it. Such prior knowledge is not necessarily available and it is not trivial to define meaningful distributions for PPRs.

In this paper, it is rather proposed to use a regularisation scheme to control the PPRs. In such a case, for a given family of models indexed by parameters $\boldsymbol{\theta} \in \Theta$, the parameter estimation consists in maximising

$$\mathcal{L}_{\Omega}\left(\boldsymbol{\theta}; \mathbf{x}, \mathbf{r}\right) = \sum_{i=1}^{n} \log\left[p\left(x_i|\boldsymbol{\theta}\right) + r_i\right] - \alpha\Omega(\mathbf{r}) \qquad (5)$$

where $\mathbf{r}$ is the vector or PPRs, α is a reinforcement meta-parameter and Ω is a penalisation function. The meta-parameter α controls the compromise between (i) fitting data using the model (large α values) and (ii) using large reinforcements to deal with data as if they were AFDs (small α values).

Using regularisation to control PPRs has several advantages: this approach remains very generic and almost no prior knowledge is required. As shown in Section 3.3, the choice of the penalisation function Ω determines the properties of the vector of PPRs, like e.g. its sparseness. Hence, it is only necessary to specify e.g. if data are expected to contain only a few *strongly* AFDs or if a lot of *weakly* AFDs are expected. This paper shows that existing statistical inference methods in machine learning can be easily adapted to use PPRs.

3.2. Generic algorithm for using pointwise probability reinforcements

In the regularised reinforced log-likelihood (5), the penalisation function Ω only depends on the PPRs in order to avoid overfitting, which could occur if Ω was also depending on the probabilities $p\left(x_i|\boldsymbol{\theta}\right)$. It also allows one to separate the optimisation of (5) in two independent steps. During the first step, the model parameters $\boldsymbol{\theta}$ are fixed to $\boldsymbol{\theta}^{\text{old}}$ and (5) is maximised only with respect to the PPRs. If Ω can be written as a sum of independent terms

$$\Omega(\mathbf{r}) = \sum_{i=1}^{n} \Omega\left(r_i\right), \qquad (6)$$

then each PPR can be optimised independently. The above condition on Ω is assumed to hold in the rest of the paper. During the second step, the PPRs are kept fixed and (5) is maximised only with respect to the model parameters $\boldsymbol{\theta}$. Since the penalisation function does not depend on the parametric probabilities, the regularisation term has no influence. The second step only works with reinforced probabilities and simply becomes

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log\left[p\left(x_i|\boldsymbol{\theta}\right) + r_i\right]. \qquad (7)$$

The two steps are further detailed in the two following subsections. Before starting the above alternate optimisation procedure, $\boldsymbol{\theta}$ has to be initialised. This can be achieved by using a method maximising the classical log-likelihood (1): the parameter estimate will be sensitive to AFDs, yet it will provide a satisfying starting point. Notice that it is important to choose a suitable value of the reinforcement meta-parameter α. A solution to this question is proposed in Section 5.

3.3. Optimisation of pointwise probability reinforcements

The expression and properties of the PPRs obtained during the first step only depend on the form of the penalisation function Ω (and not on the parametric model). Indeed, the probabilities $p\left(x_i|\boldsymbol{\theta}\right)$ are treated as fixed quantities $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ and Ω is assumed to be independent of the model parameters. In this section, L_1 and L_2 PPR regularisations are considered, which lead to sparse and non-sparse PPRs, respectively. The properties of PPRs and reinforced probabilities are also considered for a general class of penalisation functions.

3.3.1. Sparse pointwise probability reinforcements using L_1 regularisation

In linear regression, it is well known that a L_1 regularisation on the weights can be used to obtain sparse weight vectors (Efron,

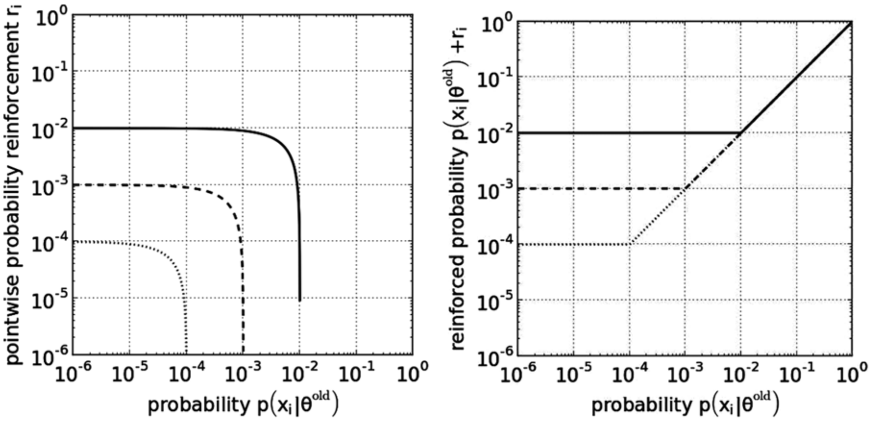


Fig. 1. PPR r_i (left) and reinforced probability $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + r_i$ (right) in terms of the probability $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ obtained using L_1 regularisation on PPRs. The reinforcement meta-parameter is $\alpha = 10^2$ (plain line), $\alpha = 10^3$ (dashed line) and $\alpha = 10^4$ (dotted line).

Hastie, Johnstone, & Tibshirani, 2004). Similarly, one can regularise PPRs using the penalisation function

$$\Omega\left(\mathbf{r}\right) = \sum_{i=1}^{n} r_i \tag{8}$$

which forces PPRs to shrink towards zero (remember that $r_i \geq 0$). The maximisation of the regularised reinforced log-likelihood leads to the Lagrangian

$$\sum_{i=1}^{n} \log\left[p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + r_i\right] - \alpha \sum_{i=1}^{n} r_i + \sum_{i=1}^{n} \beta_i r_i, \tag{9}$$

what gives the optimality condition

$$\frac{1}{p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + r_i} - \alpha + \beta_i = 0 \tag{10}$$

for each PPR r_i. When $r_i > 0$, the Lagrange multiplier β_i becomes zero and

$$r_i = \frac{1}{\alpha} - p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right). \tag{11}$$

Otherwise, when $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) > \frac{1}{\alpha}$, β_i has to be non-zero, what causes r_i to become zero. Hence, the PPR for the ith instance is

$$r_i = \max\left(\frac{1}{\alpha} - p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right), 0\right), \tag{12}$$

whereas the corresponding reinforced probability is

$$p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + r_i = \max\left(p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right), \frac{1}{\alpha}\right). \tag{13}$$

Fig. 1 shows the PPR and reinforced probability in terms of the probability for different values of the reinforcement meta-parameter α. As long as $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ remains small, the PPR is approximately equal to $\frac{1}{\alpha}$ and the reinforced probability is exactly equal to $\frac{1}{\alpha}$. Such observations are considered as potential AFDs. However, as soon as $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) \geq \frac{1}{\alpha}$, the PPR becomes zero and the reinforced probability becomes exactly equal to $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$. In such a case, the observation is no longer considered as an AFD. Interestingly, for fixed parameters $\boldsymbol{\theta}^{\text{old}}$, using a L_1 PPR regularisation is equivalent to clipping the penalised probabilities which are below $\frac{1}{\alpha}$.

In conclusion, using a L_1 regularisation leads to a very simple optimisation step for the PPRs. Moreover, the resulting PPRs are sparse and only a few of them are non-zero. Indeed, only the observations for which the probability $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ is smaller than $\frac{1}{\alpha}$ are considered as potential AFDs and reinforced accordingly. The inverse of the reinforcement meta-parameter α corresponds to the threshold applied to the reinforced probabilities.

3.3.2. *Smooth non-sparse pointwise probability reinforcements using L_2 regularisation*

A drawback of L_1 regularisation is that discontinuities may occur during the optimisation: probabilities which are reinforced at a given iteration may become unreinforced at the next iteration. Similarly to linear regression, the L_2 regularisation provides similar but smoother solutions than the L_1 regularisation. In that case, the penalisation function is

$$\Omega\left(\mathbf{r}\right) = \frac{1}{2} \sum_{i=1}^{n} r_i^2 \tag{14}$$

and has the advantage of having a zero derivative with respect to r_i when $r_i = 0$, what leads to a smoother solution. The maximisation of the regularised reinforced log-likelihood leads to the Lagrangian

$$\sum_{i=1}^{n} \log\left[p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + r_i\right] - \frac{\alpha}{2} \sum_{i=1}^{n} r_i^2 + \sum_{i=1}^{n} \beta_i r_i, \tag{15}$$

what gives for each PPR r_i the optimality condition

$$\frac{1}{p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + r_i} - \alpha r_i + \beta_i = 0. \tag{16}$$

Since $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) \geq 0$ and $\beta_i \geq 0$, it is impossible to have $r_i = 0$ and the Lagrange multiplier β_i is therefore always zero. Hence, the PPR is

$$r_i = \frac{-p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + \sqrt{p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)^2 + \frac{4}{\alpha}}}{2}, \tag{17}$$

whereas the corresponding reinforced probability is

$$p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + r_i = \frac{p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + \sqrt{p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)^2 + \frac{4}{\alpha}}}{2}. \tag{18}$$

Fig. 2 shows the PPR and reinforced probability in terms of the probability $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ for different values of the reinforcement meta-parameter α. As long as $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ remains small, the PPR and the reinforced probability are approximately equal to $\sqrt{\alpha^{-1}}$. However, as soon as $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ gets close to $\sqrt{\alpha^{-1}}$, the PPR starts decreasing towards zero and the reinforced probability tends to $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$. A comparison with Fig. 1 shows that L_2 regularisation is similar to L_1 regularisation, with smoother results and a threshold $\sqrt{\alpha^{-1}}$ instead of $\frac{1}{\alpha}$.

In conclusion, using L_2 regularisation seems to give similar results than L_1 regularisation, except that the PPRs cannot be sparse and that all observations are considered as AFDs to a certain degree. Also, the PPRs and the reinforced probabilities change more smoothly when $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ increases.

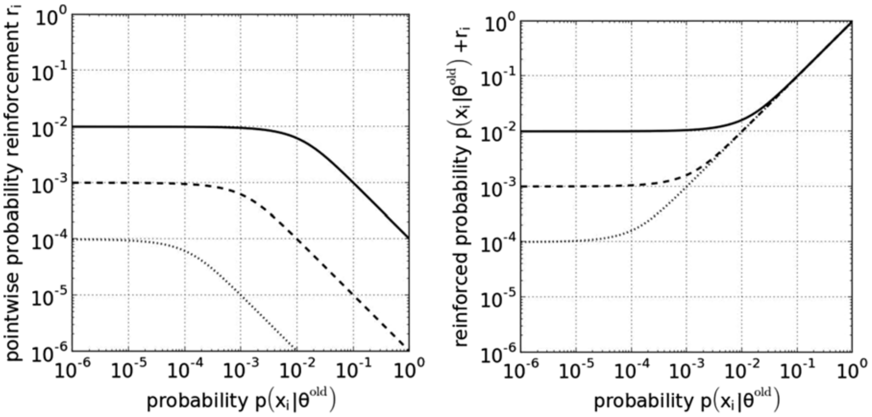


Fig. 2. PPR r_i (left) and reinforced probability $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + r_i$ (right) in terms of the probability $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ obtained using L_2 regularisation on PPRs. The reinforcement meta-parameter is $\alpha = 10^4$ (plain line), $\alpha = 10^6$ (dashed line) and $\alpha = 10^8$ (dotted line).

3.3.3. Properties of pointwise probability reinforcements and reinforced probabilities

The above derivations show that a simple expression for PPRs can be obtained using L_1 or L_2 penalisation. General results can be obtained for PPRs under reasonable requirements, e.g. that Ω is an increasing convex penalisation function which can be written as

$$\Omega\left(\mathbf{r}\right) = \sum_{i=1}^{n} \Omega\left(r_i\right) \tag{19}$$

and that the quantity $\log\left[p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + r_i\right] - \alpha\Omega\left(r_i\right)$ achieves a maximum value for a finite PPR $r_i \geq 0$ (which means that the PPR optimisation admits a solution).

In the following theorems, penalisation functions are assumed to comply with the above requirements. Also, the notation $p_i = p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ is used in order to make the developments easier to follow.

Theorem 1. *Let Ω be an increasing penalisation function of the form (19), $\boldsymbol{\theta}^{old}$ be a fixed parameter value and r_1 and r_2 be the finite optimal PPRs with respect to $\boldsymbol{\theta}^{old}$ for the observations x_1 and x_2, respectively. Then one has $r_2 \leq r_1$ if the probabilities satisfy $p\left(x_2|\boldsymbol{\theta}^{old}\right) > p\left(x_1|\boldsymbol{\theta}^{old}\right)$.*

Proof. Let us prove the theorem by showing that any PPR $r > r_1$ is suboptimal for the observation x_2. Since $p_2 > p_1$, it follows that $p_2 + r > p_2 + r_1 > p_1 + r_1$ and that $p_2 + r > p_1 + r > p_1 + r_1$. Because the logarithm is a strictly concave function, one therefore obtains the inequalities

$$\log\left[p_2 + r_1\right] > \frac{\log\left[p_2 + r\right] - \log\left[p_1 + r_1\right]}{\left(p_2 + r\right) - \left(p_1 + r_1\right)}\left(p_2 - p_1\right) + \log\left[p_1 + r_1\right] \tag{20}$$

and

$$\log\left[p_1 + r\right] > \frac{\log\left[p_2 + r\right] - \log\left[p_1 + r_1\right]}{\left(p_2 + r\right) - \left(p_1 + r_1\right)}\left(r - r_1\right) + \log\left[p_1 + r_1\right]. \tag{21}$$

Since r_1 is the optimal PPR for p_1, it also follows that

$$\log\left[p_1 + r_1\right] - \alpha\Omega\left(r_1\right) \geq \log\left[p_1 + r\right] - \alpha\Omega\left(r\right) \tag{22}$$

for any PPR r. Summing (20)–(22) eventually gives

$$\log\left[p_2 + r_1\right] - \alpha\Omega\left(r_1\right) > \log\left[p_2 + r\right] - \alpha\Omega\left(r\right), \tag{23}$$

which means that any PPR $r > r_1$ is necessarily suboptimal with respect to the probability p_2, since the PPR r_2 satisfies by definition

$$\log\left[p_2 + r_2\right] - \alpha\Omega\left(r_2\right) \geq \log\left[p_2 + r_1\right] - \alpha\Omega\left(r_1\right). \quad \square \tag{24}$$

The above theorem means that data which are more probable with respect to the parametric model are going to be less reinforced, what seems natural. Indeed, reinforcements are supposed to support unlikely observations. Notice that if there is an observation with a zero reinforcement, the reinforcements for observations with larger probabilities are also zero.

Theorem 2. *Let Ω be an increasing convex penalisation function of the form (19), $\boldsymbol{\theta}^{old}$ be a fixed parameter value and r_1 and r_2 be the finite optimal PPRs with respect to $\boldsymbol{\theta}^{old}$ for the observations x_1 and x_2, respectively. Then the reinforced probabilities are such that $p\left(x_2|\boldsymbol{\theta}^{old}\right) + r_2 \geq p\left(x_1|\boldsymbol{\theta}^{old}\right) + r_1$ if the probabilities satisfy $p\left(x_2|\boldsymbol{\theta}^{old}\right) > p\left(x_1|\boldsymbol{\theta}^{old}\right)$.*

Proof. Let us prove the theorem by considering two cases: $r_1 - (p_2 - p_1) \leq 0$ and $r_1 - (p_2 - p_1) > 0$. In the first case, since r_2 must be positive, it necessarily follows that $r_1 - (p_2 - p_1) \leq r_2$ or, by rearranging terms, that $p_2 + r_2 \geq p_1 + r_1$. In the second case, it follows from the condition $p_2 > p_1$ that $r_1 > 0$. Since r_1 is the optimal PPR for p_1, this implies that the derivative of $\log\left[p_1 + r\right] - \alpha\Omega\left(r\right)$ with respect to r is zero at $r = r_1$, i.e.

$$\frac{1}{p_1 + r_1} - \alpha\Omega'(r_1) = 0. \tag{25}$$

Moreover, the derivative of $\log\left[p_2 + r\right] - \alpha\Omega\left(r_2\right)$ at $r = r_1 - (p_2 - p_1)$ is

$$\frac{1}{p_1 + r_1} - \alpha\Omega'\left(r_1 - (p_2 - p_1)\right) \tag{26}$$

and, since $r_1 - (p_2 - p_1) < r_1$ and Ω is a convex function, it also follows that

$$\Omega'\left(r_1 - (p_2 - p_1)\right) \leq \Omega'\left(r_1\right). \tag{27}$$

Using the three above results, one can show that

$$\frac{1}{p_1 + r_1} - \alpha\Omega'\left(r_1 - (p_2 - p_1)\right) \geq 0, \tag{28}$$

i.e. that the derivative of $\log\left[p_2 + r\right] - \alpha\Omega\left(r\right)$ at $r_1 - (p_2 - p_1)$ is positive. Since this function is strictly concave in terms of r, it has only one maximum and the optimal PPR r_2 must therefore be larger or equal to this value, i.e. $r_2 \geq r_1 - (p_2 - p_1)$ or, by rearranging terms, $p_2 + r_2 \geq p_1 + r_1$. $\square$

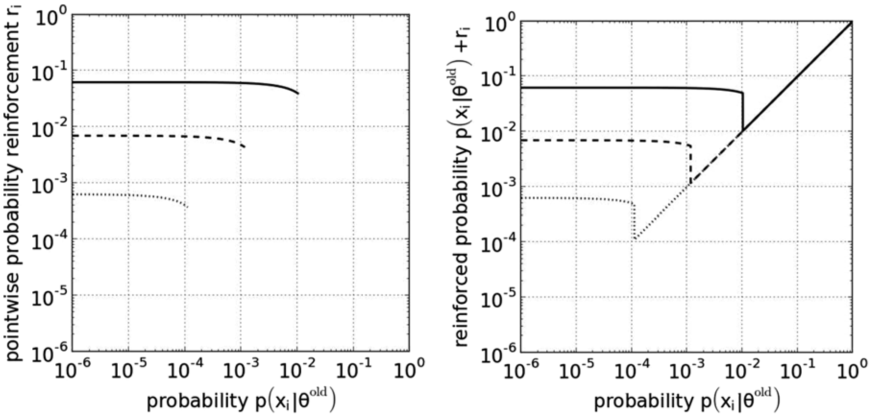


Fig. 3. PPR r_i (left) and reinforced probability $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + r_i$ (right) in terms of the probability $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ obtained using $L_{\frac{1}{2}}$ regularisation on PPRs. The reinforcement meta-parameter is $\alpha = 4$ (plain line), $\alpha = 12$ (dashed line) and $\alpha = 40$ (dotted line). Discontinuities occur at $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) \approx 10^{-4}$, $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) \approx 10^{-3}$ and $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) \approx 10^{-2}$, respectively.

The above theorem means that observations which are more probable with respect to the parametric model also correspond to larger reinforced probabilities. All other things being equal, the opposite would mean that the ordering of observations with respect to their parameterised and reinforced probabilities could be different, what seems counter-intuitive.

To illustrate the above results, let us again consider L_1 and L_2 regularisation, which use increasing and convex penalisation functions. It can be seen in Figs. 1 and 2 that the resulting PPRs and reinforced probabilities behave according to Theorems 1 and 2. A simple counter-example is $L_{\frac{1}{2}}$ regularisation, where the increasing but concave penalisation function is $\Omega(\mathbf{r}) = 2\sum_{i=1}^{n}\sqrt{r_i}$. Fig. 3 shows the PPR and reinforced probability in terms of the probability $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ for different values of the reinforcement meta-parameter α. In particular, for small values of $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$, the reinforced probability $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + r_i$ decreases when the probability $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ increases, what is a rather counter-intuitive behaviour. The PPR and the reinforced probability present a discontinuity when $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) \approx 0.65/4\alpha^2$.

3.4. Learning model parameters with pointwise probability reinforcements

The reinforced log-likelihood may be hard to maximise with respect to the model parameters. For example, if a member of the exponential family

$$p\left(x_i|\boldsymbol{\theta}\right) = h(x_i)\exp\left[\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)^T \mathbf{T}(x_i) - \psi\left(\boldsymbol{\theta}\right)\right] \tag{29}$$

is used to model data (what includes e.g. Gaussian, exponential, gamma, beta or Poisson distributions Bernardo & Smith, 2007; Bishop, 2006; DasGupta, 2011; Duda & Hart, 1973), the optimality condition becomes

$$\sum_{i=1}^{n}\frac{p\left(x_i|\boldsymbol{\theta}^{\text{new}}\right)}{p\left(x_i|\boldsymbol{\theta}^{\text{new}}\right) + r_i}\left[\boldsymbol{\eta}'\left(\boldsymbol{\theta}^{\text{new}}\right)^T \mathbf{T}(x_i) - \psi'(\boldsymbol{\theta}^{\text{new}})\right] = 0 \tag{30}$$

where $h, \boldsymbol{\eta}, \mathbf{T}$ and ψ depend on the distribution. Obviously, it will in general not be trivial to find a solution satisfying the above condition. For example, in the particular case of a univariate Gaussian distribution with unknown mean μ and known width σ, one can e.g. obtain the parameterisation (DasGupta, 2011)

$$h(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x_i^2}{2\sigma^2}}; \qquad \boldsymbol{\eta}\left(\boldsymbol{\theta}\right) = \left(\frac{\mu}{\sigma^2}\right); \tag{31}$$

$$\mathbf{T}(x_i) = \left(x_i\right); \qquad \psi\left(\boldsymbol{\theta}\right) = \frac{\mu^2}{2\sigma^2} \tag{32}$$

and eventually obtain the optimality condition

$$\sum_{i=1}^{n}\frac{\mathcal{N}\left(x_i|\mu^{\text{new}}, \sigma\right)}{\mathcal{N}\left(x_i|\mu^{\text{new}}, \sigma\right) + r_i}\left(x_i - \mu^{\text{new}}\right) = 0 \tag{33}$$

where μ^{new} cannot be isolated.

Rather than directly optimising the reinforced log-likelihood, this paper proposes to indirectly maximise it by iteratively (i) finding a close lower bound to the reinforced log-likelihood and (ii) maximising this bound with respect to $\boldsymbol{\theta}$. This procedure is similar to the widely used expectation maximisation (EM) algorithm (Dempster, Laird, & Rubin, 1977). The following theorem shows that it is easily possible to find such a lower bound for any parametric model.

Theorem 3. *Let $\boldsymbol{\theta}^{old}$ be the current estimate of the model parameters and, for each observation x_i, let r_i be the optimal PPR with respect to $\boldsymbol{\theta}^{old}$. If one defines the observation weight*

$$w_i = \frac{p\left(x_i|\boldsymbol{\theta}^{old}\right)}{p\left(x_i|\boldsymbol{\theta}^{old}\right) + r_i}, \tag{34}$$

then the functional

$$\sum_{i=1}^{n}\left[w_i \log\frac{p\left(x_i|\boldsymbol{\theta}\right)}{p\left(x_i|\boldsymbol{\theta}^{old}\right)} + \log\left[p\left(x_i|\boldsymbol{\theta}^{old}\right) + r_i\right]\right] \tag{35}$$

is a lower bound to the reinforced log-likelihood

$$\sum_{i=1}^{n}\log\left[p\left(x_i|\boldsymbol{\theta}\right) + r_i\right]. \tag{36}$$

Moreover, (36) and (35) are tangent at $\boldsymbol{\theta} = \boldsymbol{\theta}^{old}$.

Proof. When $\boldsymbol{\theta} = \boldsymbol{\theta}^{\text{old}}$, the value of (35) and (36) is

$$\sum_{i=1}^{n}\log\left[p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + r_i\right], \tag{37}$$

whereas their derivative with respect to model parameters is

$$\sum_{i=1}^{n}\frac{1}{p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right) + r_i}\frac{\delta p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)}{\delta\boldsymbol{\theta}}. \tag{38}$$

Since their value and derivative are identical in $\boldsymbol{\theta} = \boldsymbol{\theta}^{\text{old}}$, it follows that (35) and (36) are tangent at that point. Let us now prove that (35) is a lower bound to (36) by considering their terms. Indeed, if each ith term of (36) is lower bounded by the ith term of (35), (35) is necessarily a lower bound to (36). Let us first rewrite the inequality

$$\log\left[p\left(x_i|\boldsymbol{\theta}\right)+r_i\right]\geq w_i\log\frac{p\left(x_i|\boldsymbol{\theta}\right)}{p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)}+\log\left[p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)+r_i\right] \quad (39)$$

as

$$\left[p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)+r_i\right]\log\left[\frac{p\left(x_i|\boldsymbol{\theta}\right)+r_i}{p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)+r_i}\right] \geq p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)\log\frac{p\left(x_i|\boldsymbol{\theta}\right)}{p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)}. \quad (40)$$

When $r_i = 0$, it is easily shown that both sides of the inequality are equal, what is natural since $w_i = 1$ in such a case. Hence, since r_i is always positive, it is sufficient to show that the derivative of the left side with respect to r_i is always larger than the derivative of the right side to ensure that the inequality (39) is verified for any $r_i \geq 0$. This condition can be written as

$$\log\left[\frac{p\left(x_i|\boldsymbol{\theta}\right)+r_i}{p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)+r_i}\right]+\frac{p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)+r_i}{p\left(x_i|\boldsymbol{\theta}\right)+r_i}-1\geq 0. \quad (41)$$

Using the standard logarithm inequality $\log x \geq \frac{x-1}{x}$, one can show that

$$\log\left[\frac{p\left(x_i|\boldsymbol{\theta}\right)+r_i}{p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)+r_i}\right]\geq\frac{p\left(x_i|\boldsymbol{\theta}\right)-p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)}{p\left(x_i|\boldsymbol{\theta}\right)+r_i} \quad (42)$$

and it follows that

$$\log\left[\frac{p\left(x_i|\boldsymbol{\theta}\right)+r_i}{p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)+r_i}\right]+\frac{p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)+r_i}{p\left(x_i|\boldsymbol{\theta}\right)+r_i}-1 \geq\frac{p\left(x_i|\boldsymbol{\theta}\right)+r_i}{p\left(x_i|\boldsymbol{\theta}\right)+r_i}-1=0, \quad (43)$$

what proves the inequality (41) and concludes the proof. □

Based on the above theorem, a maximisation step can easily be found. Indeed, since (35) is a lower bound to the reinforced log-likelihood and both are tangent at $\boldsymbol{\theta} = \boldsymbol{\theta}^{\text{old}}$, maximising the former will necessarily increase the latter with respect to its value in $\boldsymbol{\theta}^{\text{old}}$. Hence, the approximate maximisation step (with respect to the model parameters $\boldsymbol{\theta}$) of the proposed algorithm consists in maximising the weighted log-likelihood

$$\mathcal{L}_{\mathbf{w}}\left(\boldsymbol{\theta};\mathbf{x}\right)=\sum_{i=1}^{n}w_i\log p\left(x_i|\boldsymbol{\theta}\right) \quad (44)$$

where weights are computed using (34) and the current estimate of the model parameters $\boldsymbol{\theta}^{\text{old}}$. The advantage of this maximisation step is that weighted log-likelihoods are typically much easier to maximise than reinforced log-likelihoods. For example, in the case of the above Gaussian distribution with known width σ, one obtains the optimality condition

$$\sum_{i=1}^{n}w_i\left(x_i-\mu^{\text{new}}\right)=0 \quad (45)$$

where

$$w_i=\frac{\mathcal{N}\left(x_i|\mu^{\text{old}},\sigma\right)}{\mathcal{N}\left(x_i|\mu^{\text{old}},\sigma\right)+r_i}. \quad (46)$$

Hence, the mean of the Gaussian is estimated by the weighed sample mean

$$\mu^{\text{new}}=\frac{\sum_{i=1}^{n}w_ix_i}{\sum_{i=1}^{n}w_i}. \quad (47)$$

Interestingly, the weighted log-likelihood is often used in the literature to reduce the impact of some of the data (Hu & Zidek, 2002). Using L_1 regularisation, the proposed approach is similar to the trimmed likelihood approach (Cheng & Biswas, 2008; Hadi & Luceo, 1997; Neykov, Filzmoser, Dimova, & Neytchev, 2007), where only a subset of the observations are used to compute the log-likelihood.

As discussed in Section 3.2, the maximisation step does not depend on the penalisation function Ω. Fig. 4 shows examples of lower-bounded reinforced log-likelihoods for 50 observations drawn from a univariate Gaussian distribution with mean $\mu = 0$ and width $\sigma = 1$. The PPRs r_i are computed using the current estimates $\mu^{\text{old}} = 0.2$ and $\sigma^{\text{old}} = 1.3$. L_1 regularisation is used in Fig. 4(a) and (b), whereas L_2 regularisation is used in Fig. 4(c) and (d). The log-likelihoods are computed for different values of the mean μ in Fig. 4(a) and (c) and different values of the width σ in Fig. 4(b) and (d). Reinforced log-likelihoods and lower bounds are tangent at $\mu = \mu^{\text{old}}$ and $\sigma = \sigma^{\text{old}}$, in accordance with Theorem 3.

Notice that at optimum, the solution $\boldsymbol{\theta}^*$ of existing weighted maximum likelihood approaches can be seen as reinforced maximum likelihood solutions where the equivalent PPR for the ith observation would be

$$r_i=\frac{1-w_i}{w_i}p\left(x_i|\boldsymbol{\theta}^*\right), \quad (48)$$

which follows from the inversion of (34) defining weights.

3.5. Weights and degrees of abnormality

Aside from its interesting properties for the optimisation of the model parameters, the weighted log-likelihood (44) also provides us a useful tool to analyse data. Indeed, when an optimum is reached, the weighted log-likelihood and the reinforced log-likelihood are equal. In such a case, the weight $w_i \in [0, 1]$ measures the contribution of the likelihood of the ith observation to these quantities. The quantity $a_i = 1 - w_i \in [0, 1]$ can be interpreted as an *abnormality degree* and may be easier to handle than the PPR r_i that belongs to $[0, \infty[$. When the PPR of an instance gets close to zero, its weight approaches one and its abnormality degree becomes zero. On the contrary, when the PPR increases, the weight tends to zero and the abnormality degree tends to one. In other words, data whose probability is highly reinforced (because they appear to be AFDs) are characterised by small weights and large abnormality degrees.

Fig. 5 shows the weight w_i in terms of the parametric probability $p\left(x_i|\theta\right)$ for L_1 and L_2 regularisation. The weight is close to one for large probabilities, but it decreases quickly as the probability decreases. On the contrary, notice that the abnormality degree a_i would be close to one for small probabilities, then decrease quickly as the probability increases.

4. Supervised and unsupervised inference using pointwise probability reinforcements

This section adapts several standard statistical inference methods to reinforce them with PPRs: linear regression, kernel ridge regression (a.k.a. least-square support vector machines), logistic

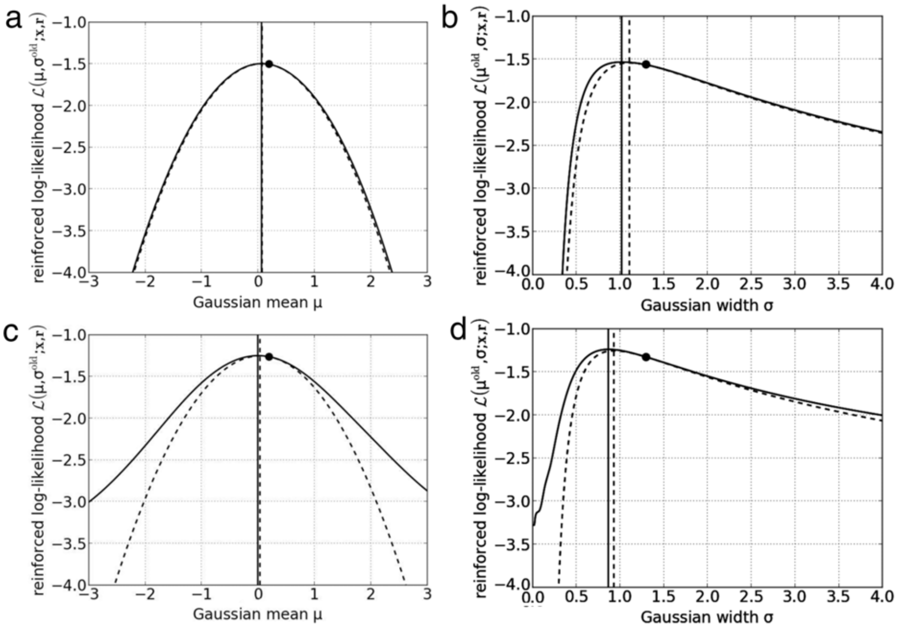


Fig. 4. Examples of tangent lower bounds (dashed line) for the reinforced log-likelihood (plain line) in a Gaussian case. The reinforced probabilities are $\mathcal{N}\left(x_i|\mu, \sigma^{\text{old}}\right) + r_i$ (left) and $\mathcal{N}\left(x_i|\mu^{\text{old}}, \sigma\right) + r_i$ (right), where PPRs are computed using either L_1 (up) or L_2 (down) regularisation. Vertical lines indicate the position of the maximum for the reinforced log-likelihood (plain line) and its lower bound (dashed line). Dots indicate points of tangency.

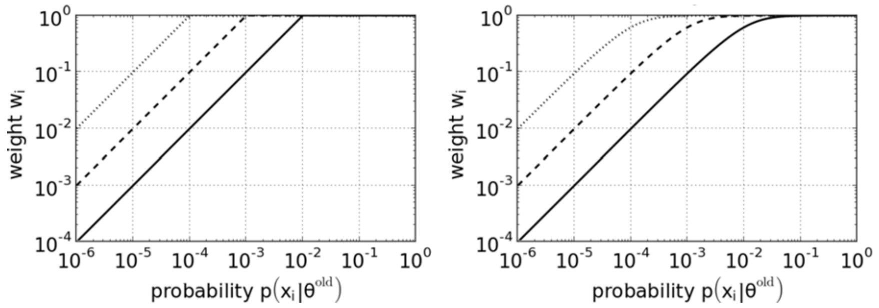


Fig. 5. Observation weight w_i in terms of the probability $p\,(x_i|\theta)$ for L_1 (left) and L_2 (right) regularisation. The L_1 reinforcement meta-parameter is $\alpha = 10^2$ (plain line), $\alpha = 10^3$ (dashed line) and $\alpha = 10^4$ (dotted line). The L_2 reinforcement meta-parameter is $\alpha = 10^4$ (plain line), $\alpha = 10^6$ (dashed line) and $\alpha = 10^8$ (dotted line).

regression and principal component analysis. These four techniques tackle regression, classification and projection problems, what shows that PPRs allow one to easily deal with AFDs in various supervised or unsupervised contexts. For each method, experiments target outliers, since outliers are commonly recognised as harmful in the above applications.

4.1. Reinforced linear regression

Linear regression consists in fitting a linear prediction model $f(x_i) = \sum_{j=1}^{d} \beta_j x_{ij} + \beta_0$ to observed target values, where x_{ij} is the value of the jth feature for the ith observation and d is the dimensionality of data. Under the assumption that a Gaussian noise pollutes the observations, the maximum likelihood solution is given by the well-known ordinary least squares (OLS) estimator

$$\boldsymbol{\beta}_{\text{OLS}} = \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{y} \tag{49}$$

where $\tilde{\mathbf{X}}$ is the $n \times (1+d)$ matrix of data with an additional column of ones and $\mathbf{y}$ is the vector of target values. For small datasets, the above estimator is quite sensitive to outliers (Beckman & Cook, 1983; Cook, 1979; Hadi & Simonoff, 1993), but it can easily be reinforced using PPRs. As discussed in Section 3.4, the model parameters optimisation step can be achieved by maximising a weighted log-likelihood $\mathcal{L}_{\mathbf{w}}\,(\boldsymbol{\beta}, \sigma; \mathbf{x}, \mathbf{y})$, which becomes

$$\frac{1}{2}\log\left[2\pi\sigma^2\right]\sum_{i=1}^{n} w_i + \frac{1}{2\sigma^2}\left(\tilde{\mathbf{X}}\boldsymbol{\beta} - \mathbf{y}\right)^T\mathbf{W}\left(\tilde{\mathbf{X}}\boldsymbol{\beta} - \mathbf{y}\right) \tag{50}$$

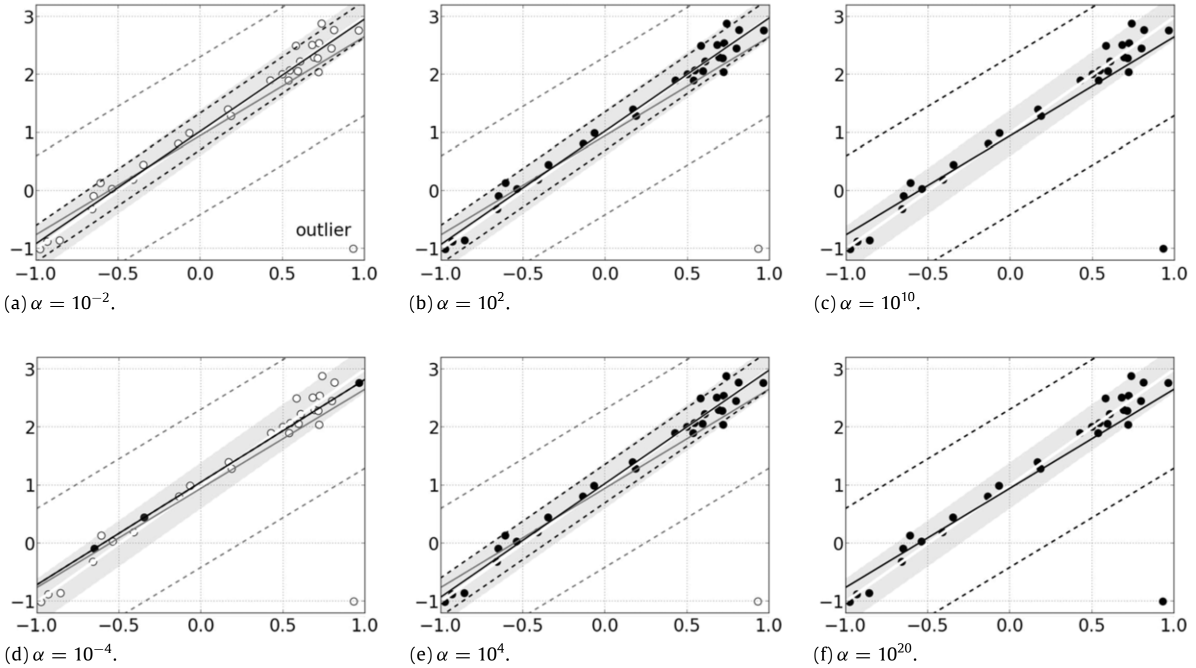
where σ is the Gaussian noise variance and $\mathbf{W}$ is a diagonal weighting matrix whose diagonal terms are $W_{ii} = w_i$. The solution maximising the above log-likelihood is similar to weighted least squares (WLS) estimator, except that the noise variance has to be also estimated. Indeed, the probabilities $p(y_i|x_i, \boldsymbol{\beta}, \sigma)$ must be estimated in order to obtain PPRs. The estimates are

$$\boldsymbol{\beta}_{\text{PPR}} = \left(\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{W}\mathbf{y} \tag{51}$$

and

$$\sigma^2_{\text{PPR}} = \frac{\sum_{i=1}^{n} w_i\left(f(x_i) - y_i\right)^2}{\sum_{i=1}^{n} w_i}. \tag{52}$$

When L_1 regularisation is used, reinforced linear regression is similar to least trimmed squares (Rousseeuw, 1984; Ruppert & Carroll, 1980), which may be expensive for large datasets (Rousseeuw & Van Driessen, 2006).



(a) $\alpha = 10^{-2}$.

(b) $\alpha = 10^{2}$.

(c) $\alpha = 10^{10}$.

(d) $\alpha = 10^{-4}$.

(e) $\alpha = 10^{4}$.

(f) $\alpha = 10^{20}$.

Fig. 6. Results obtained by standard linear regression (grey line) and reinforced linear regression (black line). The 95% confidence interval associated with the true function (white line) is shown by the grey-shaded area. PPRs are computed using L_1 (upper row) and L_2 (lower row) regularisation. The reinforcement meta-parameter values are indicated below each subfigure. The 30 data are shown by circles whose darkness is proportional to their respective weights. Dashed lines delimit 95% confidence intervals for prediction of each model.

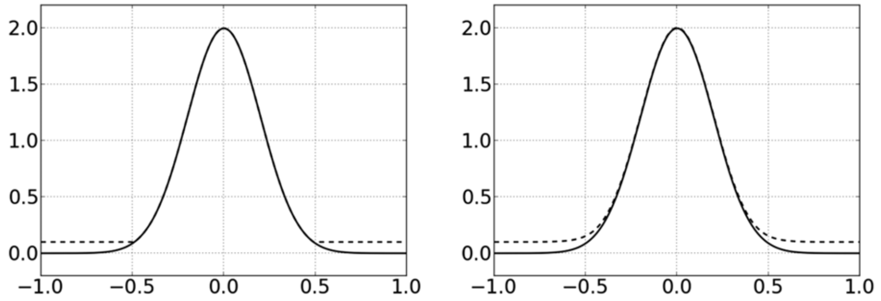


Fig. 7. Comparison of the probability $\mathcal{N}(\epsilon_i|0, 0.2)$ (plain line) and the reinforced probability $\mathcal{N}(\epsilon_i|0, 0.2) + r_i$ (dashed line) of the residual $\epsilon_i = y_i - f(x_i)$ obtained using L_1 (left) and L_2 (right) regularisation. The reinforcement meta-parameter values are $\alpha_1 = 25$ and $\alpha_2 = 625$, respectively.

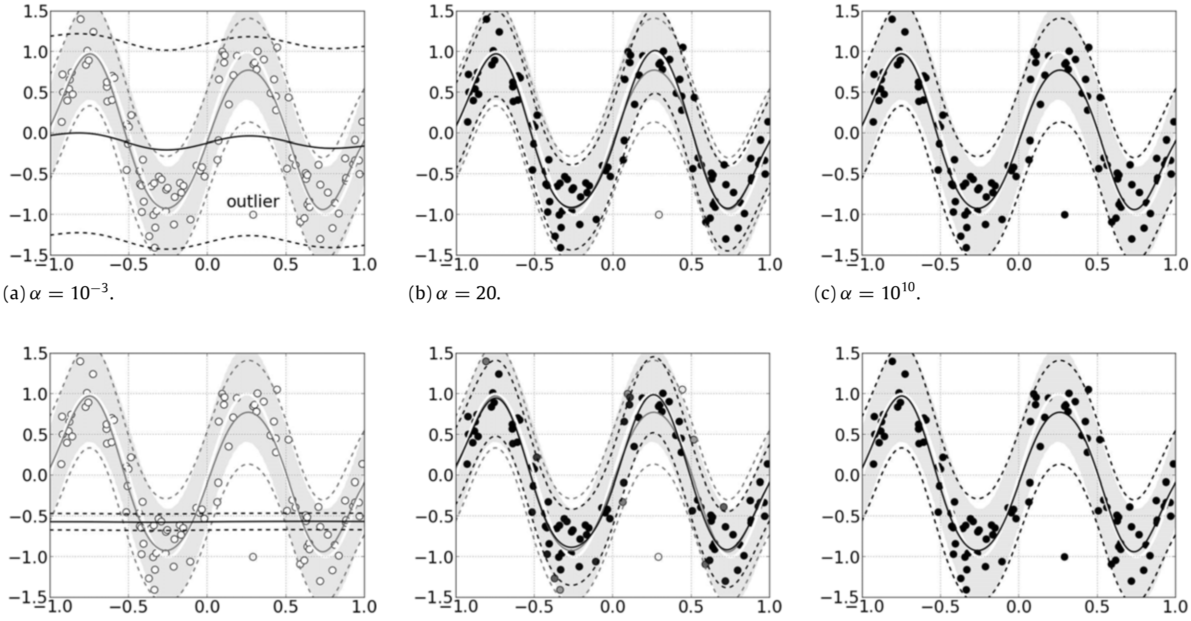
Fig. 6 shows results obtained using reinforced linear regression with L_1 and L_2 regularisation on the PPRs. Notice that the reinforced linear regression estimates an almost zero noise variance in Fig. 6(d). Standard and reinforced solutions are superimposed in Fig. 6(c) and (f). 29 data are generated from a unidimensional linear model with $\boldsymbol{\beta} = [2, 1]^T$ and $\sigma = 0.2$. Moreover, one outlier is added in order to interfere with inference. As a result, Fig. 6 shows that standard linear regression is biased and that its confidence interval for prediction is quite wide. On the one hand, Fig. 6(a) and (d) show that when the reinforcement meta-parameter α is too small, the outlier is not detected. For L_1 penalisation, all weights are identical, whereas they seem quite arbitrary for L_2 regularisation. Moreover, the confidence interval for the prediction of the reinforced linear regression is very narrow in that latter case. On the other hand, Fig. 6(c) and (f) show that when α is very large, the reinforced linear regression obtains results which are similar to standard linear regression results, since PPRs are forced to take very small values. A good compromise is obtained in Fig. 6(b) and (e) where the intermediate value of α allows PPRs to detect the outlier. Hence, the model is freed from its influence, what results in a more reliable model and improved 95% confidence intervals for predictions. Section 5 shows how to find an intermediate value of α corresponding to this compromise.

In order to illustrate the effect of the reinforcement of probabilities, let us consider the probability $\mathcal{N}(\epsilon_i|0, 0.2)$ of the residual $\epsilon_i = y_i - f(x_i)$. Fig. 7 shows the reinforced probability $\mathcal{N}(\epsilon_i|0, 0.2) + r_i$ obtained using L_1 and L_2 regularisation; the effect of the probability reinforcement is to clip the Gaussian distribution in the tails. Indeed, when $\mathcal{N}(\epsilon_i|0, 0.2)$ gets too small, it is replaced by $\frac{1}{\alpha}$ for L_1 regularisation and $\sqrt{\alpha^{-1}}$ for L_2 regularisation. The reinforced probability tends to $\frac{1}{\alpha_1} = 0.04$ and $\sqrt{\alpha_2^{-1}} = 0.04$ in the tails.

The reinforced linear regression allows us to deal with outliers, but it can also easily be adapted to penalise large feature weights β_j. Indeed, if a penalisation $\frac{\gamma}{2}\|\boldsymbol{\beta}\|^2$ with a regularisation meta-parameter γ is added to the reinforced log-likelihood, one obtains a reinforced ridge regression (Hoerl & Kennard, 1970) where the weights are estimated by

$$\boldsymbol{\beta}_{\text{PPR}} = \left(\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}} + \gamma\mathbf{I}_d\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{W}\mathbf{y} \tag{53}$$

where $\mathbf{I}_d$ is the $d \times d$ identity matrix. Interestingly, it can be seen that both regularisations coexist without problem in the above solution. Each regularisation takes independently care of one problem, what allows to readily separate outlier control and complexity control in a theoretically sound way.



(a) $\alpha = 10^{-3}$.

(b) $\alpha = 20$.

(c) $\alpha = 10^{10}$.

(d) $\alpha = 10^{-6}$.

(e) $\alpha = 40$.

(f) $\alpha = 10^{20}$.

Fig. 8. Results obtained by standard kernel ridge regression (grey line) and reinforced kernel ridge regression (black line). The 95% confidence interval associated with the true function (white line) is shown by the grey-shaded area. PPRs are computed using L_1 (upper row) and L_2 (lower row) regularisation. The reinforcement meta-parameter values are indicated below each subfigure. The 30 data are shown by circles whose darkness is proportional to their respective weights. Dashed lines delimit 95% confidence intervals.

4.2. Reinforced kernel ridge regression

Kernel ridge regression (Saunders, Gammerman, & Vovk, 1998) (also called least squares support vector machine (Suykens, Van Gestel, De Brabanter, De Moor, & Vandewalle, 2002)) is an extension of ridge regression where data are first mapped in a feature space. This kernel-based non-linear regression method can be formulated as a maximum likelihood problem. Indeed, ridge regression corresponds to assuming that the n errors $\epsilon_i = y_i - f(x_i)$ follow a Gaussian distribution $\mathcal{N}\left(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n\right)$ and that the m weights in the feature space have a Gaussian prior $\mathcal{N}\left(0, \sigma_\beta^2 \mathbf{I}_m\right)$, excluding the bias β_0. In such a case, it turns out that the prediction function is

$$f(x_i) = \sum_{j=1}^{n} \alpha_j k(x_i, x_j) + \beta_0 \tag{54}$$

where $\alpha_j > 0$ are dual weights and the kernel function k computes the dot product between two observations in the feature space (Muller, Mika, Ratsch, Tsuda, & Scholkopf, 2001). If one introduces the meta-parameter $\gamma = \sigma_\beta^2 / \sigma_\epsilon^2$ which controls the compromise between errors and model complexity and whose value is chosen using e.g. cross-validation, the parameter estimate of $\boldsymbol{\alpha}$ and β_0 is the solution of the linear system

$$\begin{pmatrix} 0 & \mathbf{1}_n^T \\ \mathbf{1}_n & K + \frac{1}{\gamma}\mathbf{I}_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{y} \end{pmatrix} \tag{55}$$

where $\mathbf{K}$ is the Gram matrix such that $K_{ij} = k(x_i, x_j)$ and $\mathbf{1}_n$ is a n-element vector of 1's. Moreover, the standard deviation σ_ϵ can be estimated as

$$\sigma_\epsilon^2 = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2. \tag{56}$$

The above kernel ridge regression can easily be reinforced. Indeed, the weighted maximum likelihood problem solved in the parameter optimisation step is equivalent to a weighted kernel ridge regression (Jiyan et al., 2011; Liu et al., 2011; Suykens, De Brabanter et al., 2002; Wen et al., 2010). It follows that the parameter estimate of $\boldsymbol{\alpha}$ and β_0 is the solution of the linear system

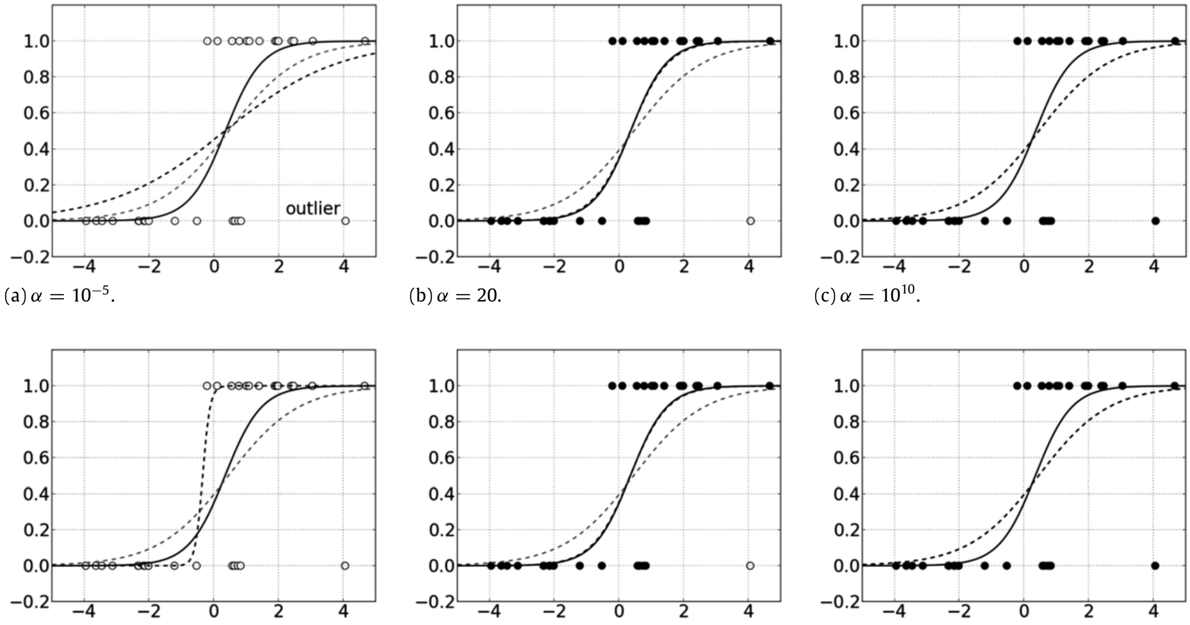
$$\begin{pmatrix} 0 & \mathbf{1}_n^T \\ \mathbf{1}_n & K + \frac{1}{\gamma}\mathbf{W}^{-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{y} \end{pmatrix}, \tag{57}$$

whereas the standard deviation σ_ϵ can be estimated as

$$\sigma_\epsilon^2 = \frac{\sum_{i=1}^{n} w_i \epsilon_i^2}{\sum_{i=1}^{n} w_i}. \tag{58}$$

The solutions in (55) and (57) are almost identical. The only difference is that the identity matrix $\mathbf{I}_n$ in (55) is replaced by the inverse of the weighting matrix $\mathbf{W}$ in (57). Since outliers correspond to large diagonal entries in $\mathbf{W}^{-1}$, the complexity control is less affected by them. Hence, the outlier control allows reducing the impact of outliers on complexity control.

Fig. 8 shows results obtained using reinforced kernel ridge regression with L_1 and L_2 regularisation on the PPRs. Standard and reinforced solutions are superimposed in Fig. 8(c) and (f). 29 data are generated from a unidimensional sinus polluted by a Gaussian noise with $\sigma = 0.1$. Moreover, one outlier is added in order to interfere with inference. For both the standard and the reinforced kernel ridge regression, the value of the meta-parameter is $\gamma = 10$ and the Gaussian kernel $k(x, z) = \exp \|x - z\|^2 / 0.1$ is used. The result of standard kernel ridge regression appears to be biased by the outlier. For small α values, Fig. 8(a) and (d) show that irrelevant models are obtained, because the $K\boldsymbol{\alpha}$ term is negligible with respect to the term $\frac{1}{\gamma}\mathbf{W}^{-1}\boldsymbol{\alpha}$ in the linear system (57). Fig. 8(c) and (f) show that reinforced kernel ridge regression performs similarly to standard kernel ridge regression when α is very large. A good compromise is obtained in Fig. 8(b) and (e) where the intermediate value of α allows PPRs to detect outliers. Indeed, the outlier



(a) $\alpha = 10^{-5}$.

(b) $\alpha = 20$.

(c) $\alpha = 10^{10}$.

(d) $\alpha = 10^{-5}$.

(e) $\alpha = 400$.

(f) $\alpha = 10^{10}$.

Fig. 9. Results obtained by standard logistic regression (grey dashed line) and reinforced logistic regression (black dashed line) from polluted data, with respect to the result obtained by standard logistic regression from clean data (black plain line). PPRs are computed using L_1 (upper row) and L_2 (lower row) regularisation. The reinforcement meta-parameter values are indicated below each subfigure. The 30 data are shown by circles whose darkness is proportional to their weights. The observed (noisy) labels are indicated by the y-value of each point (0 or 1).

is clearly detected and the resulting non-linear regression model is freed from its influence. Moreover, 95% confidence intervals for predictions are also more reliable using PPRs. Section 5 shows how to find an intermediate value of α for a good compromise.

4.3. Reinforced logistic regression

Logistic regression is a standard classification model which linearly discriminates between two classes (0 and 1 here). Conditional class probabilities for this model are obtained using

$$p\left(Y_i = 1|X_i = x_i\right) = 1 - p\left(Y_i = 0|X_i = x_i\right) = \frac{1}{1 + e^{-\sum_{j=1}^{d} \beta_j x_{ij} - \beta_0}} \qquad (59)$$

where Y_i is the class of the ith observation. Using the iterative reweighted least squares (IRLS) algorithm (Bishop, 2006), logistic regression can be efficiently performed. This quasi-Newton approach method consists in using the estimate

$$\boldsymbol{\beta}_{\text{IRLS}} = \left(\tilde{\mathbf{X}}^T \mathbf{R} \tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}^T \mathbf{R} \mathbf{z} \qquad (60)$$

iteratively, where $\tilde{\mathbf{X}}$ is the $n \times (1 + d)$ matrix of data with an additional column of ones, $\mathbf{R}$ is a diagonal matrix whose diagonal terms are $R_{ii} = \sigma_i (1 - \sigma_i)$ with $\sigma_i = p\left(Y_i = 1|X_i = x_i\right)$ and $\mathbf{z}$ is a vector of altered targets

$$\mathbf{z} = \tilde{\mathbf{X}} \boldsymbol{\beta} - \mathbf{R}^{-1} (\boldsymbol{\sigma} - \mathbf{y}). \qquad (61)$$

Logistic regression is sensitive to outliers (Rousseeuw & Christmann, 2003), but it can be reinforced using PPRs. Since the model parameter optimisation step is in fact a weighted logistic regression (Rousseeuw & Christmann, 2003; Simeckova, 2005), it can also be performed by an IRLS-like algorithm. The only modification is that the iterative update becomes

$$\boldsymbol{\beta}_{\text{IRLS}} = \left(\tilde{\mathbf{X}}^T \mathbf{W} \mathbf{R} \tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}^T \mathbf{W} \mathbf{R} \mathbf{z}. \qquad (62)$$

Fig. 9 shows results obtained using reinforced logistic regression with L_1 and L_2 regularisation on the PPRs. The reinforced solution and the standard solution obtained from clean data are superimposed in Fig. 9(b) and (e). The standard solution obtained from polluted data and the reinforced solution are superimposed in Fig. 9(c) and (f). 30 data are generated from two classes with Gaussian distributions $\mathcal{N}(\mu = \pm 2, \sigma = 1.7)$. In order to introduce an outlier, the label of one observation from class 1 is flipped, i.e. a labelling error is introduced which alters the result of standard logistic regression. On the one hand, Fig. 9(a) and (d) show that when the reinforcement meta-parameter α is too small, the outlier is not detected. On the other hand, Fig. 9(c) and (f) show that when α is very large, the reinforced logistic regression obtains results which are similar to standard logistic regression results with polluted data, since PPRs are forced to take very small values. A good compromise is obtained in Fig. 9(b) and (e) where the reinforced logistic regression produces a model which is very close to the model obtained by standard logistic regression with no labelling error. Section 5 shows how to find an intermediate value of α which allows a good compromise.

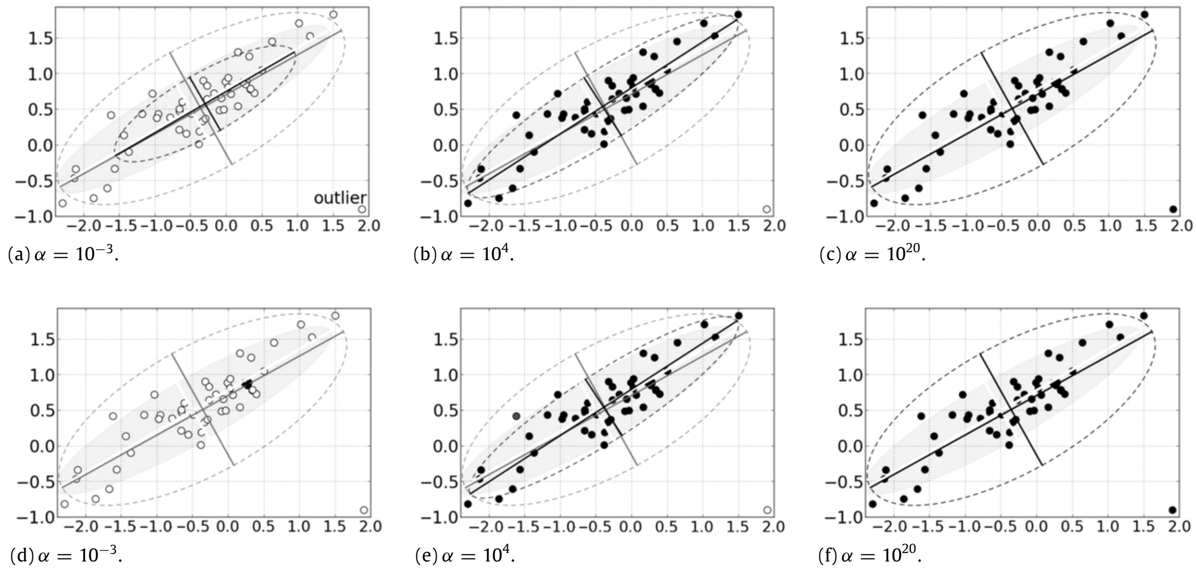
4.4. Reinforced principal component analysis

Principal component analysis (PCA) finds the q principal (or maximum variance) axes of a data cloud. This unsupervised procedure projects data onto a smaller dimensional space, while keeping the most of the feature variance. PCA can be cast as a probabilistic method (Tipping & Bishop, 1999) by assuming that (i) data are generated by q hidden independent Gaussian sources Z with distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_q)$ and (ii) that one only observes d features X whose conditional distribution is

$$p\left(X = x|Z = z, \boldsymbol{\mu}, \sigma\right) = \mathcal{N}\left(x|\mathbf{A}z + \boldsymbol{\mu}, \sigma^2 \mathbf{I}_d\right) \qquad (63)$$

where $\mathbf{A}$ is $d \times q$ linear transformation matrix, μ is a d-dimensional translation vector and σ is the noise standard deviation. Tipping and Bishop (1999) show that the observed features have a marginal distribution

$$p\left(X = x|\boldsymbol{\mu}, \sigma\right) = \mathcal{N}\left(x|\boldsymbol{\mu}, \mathbf{C}\right) \qquad (64)$$



(a) $\alpha = 10^{-3}$.

(b) $\alpha = 10^{4}$.

(c) $\alpha = 10^{20}$.

(d) $\alpha = 10^{-3}$.

(e) $\alpha = 10^{4}$.

(f) $\alpha = 10^{20}$.

Fig. 10. Axes obtained by standard PCA (grey lines) and reinforced PCA (black lines), with respect to the true axes of the hidden data model (white lines) for which the grey-shaded area shows the true 95% confidence region. PPRs are computed using L_1 (upper row) and L_2 (lower row) regularisation. The reinforcement meta-parameter values are indicated below each subfigure. The 50 data are shown by circles whose darkness is proportional to their respective weights. Dashed lines are level curves of the Gaussian distribution which delimit the 95% confidence region for each model, except in (d) where the reinforced PCA estimates an almost zero variance in the second principal axis direction. In (c) and (f), standard and reinforced solutions are superimposed.

where $\mathbf{C} = \mathbf{A}\mathbf{A}^T + \sigma^2 \mathbf{I}_d$. Hence, the data log-likelihood is

$$\mathcal{L}(\mathbf{A}, \sigma; \mathbf{x}) = -\frac{n}{2}\left[d \log[2\pi] + \log|\mathbf{C}| + \mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{S}\right)\right] \tag{65}$$

where the sample covariance matrix $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \boldsymbol{\mu})(x_i - \boldsymbol{\mu})^T$ is computed using the sample mean $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The maximum likelihood solution is

$$\mathbf{A}_{\mathrm{ML}} = \mathbf{U}_q\left(\boldsymbol{\Lambda}_q - \sigma^2 \mathbf{I}_q\right)^{\frac{1}{2}} \tag{66}$$

where $\mathbf{U}_q$ contains the q principal eigenvectors of $\mathbf{S}$ as columns and $\boldsymbol{\Lambda}_q$ is a diagonal matrix containing the q corresponding eigenvalues (Tipping & Bishop, 1999). Moreover, the maximum likelihood estimator of σ when $\mathbf{A} = \mathbf{A}_{\mathrm{ML}}$ is given for $d > q$ by

$$\sigma^2_{\mathrm{ML}} = \frac{1}{d - q}\sum_{i=q+1}^{d} \lambda_i. \tag{67}$$

When $d = q$, data are reconstructed with no error and σ_{ML} is zero. PCA is known to be sensitive to outliers (Archambeau et al., 2006; Daszykowski et al., 2007; Filzmoser, Maronna, & Werner, 2008; Hubert, Rousseeuw, & Verdonck, 2009; Stanimirova, Walczak, Massart, & Simeonov, 2004; Xu & Yuille, 1995), but this method can be easily reinforced. Indeed, it turns out that the parameter optimisation step is a weighted PCA (Fan et al., 2011; Huber, 1981), which simply consists in using (66) and (67) with the eigenvectors and eigenvalues of the weighted sample covariance matrix

$$\mathbf{S} = \frac{\sum_{i=1}^{n} w_i (x_i - \boldsymbol{\mu})(x_i - \boldsymbol{\mu})^T}{\sum_{i=1}^{n} w_i} \tag{68}$$

where $\boldsymbol{\mu}$ is the weighted sample mean

$$\boldsymbol{\mu} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}. \tag{69}$$

Similarly to the other reinforced methods, the weights are obtained using the definition (34) and the marginal probabilities given by (64).

Fig. 10 shows the results obtained using reinforced PCA with L_1 and L_2 regularisation on the PPRs. 49 data are (i) generated from a two-dimensional isotropic Gaussian distribution with mean $[0, 0]^T$ and covariance matrix $\mathbf{I}_2$ and (ii) transformed using the linear transformation matrix

$$\mathbf{A} = \begin{pmatrix} \cos\frac{\pi}{6} & -\sin\frac{\pi}{6} \\ \sin\frac{\pi}{6} & \cos\frac{\pi}{6} \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0.2 \end{pmatrix} = \begin{pmatrix} 0.87 & -0.1 \\ 0.5 & 0.87 \end{pmatrix} \tag{70}$$

and the translation matrix $\mu = [-0.5, 0.5]^T$. Moreover, one outlier is added in order to interfere with inference. As a result, the axes given by standard PCA are slightly rotated and stretched. On the one hand, Fig. 10(a) and (d) show that when the reinforcement meta-parameter α is too small, unconvincing results are obtained because PPRs are free to take large values. On the other hand, Fig. 10(c) and (f) show that when α is very large, our approach obtains results which are similar to standard PCA results, since PPRs are forced to take very small values. A good compromise is obtained in Fig. 10(b) and (e) with an intermediate value of α. Section 5 shows how to find such an intermediate value of α.

5. Choice of the reinforcement meta-parameter

As mentioned in Section 3.2, the reinforcement meta-parameter α determines the amount of reinforcement which is allowed. As illustrated by the results of the various reinforced methods presented in Section 4, small values of α lead to large PPRs, whereas large values of α lead to small PPRs. In the former case, the parametric model is insensitive to AFDs but may poorly fit data, since all of them are seen as AFDs with large PPRs. In the latter case, none of the data can be seen as an AFD and the parametric model is given by the standard maximum likelihood solution, which is sensitive to AFDs. In conclusion, it is important to choose a good intermediate value of α where only a few data can be considered as AFDs with large PPRs, what results in model parameters which make sense

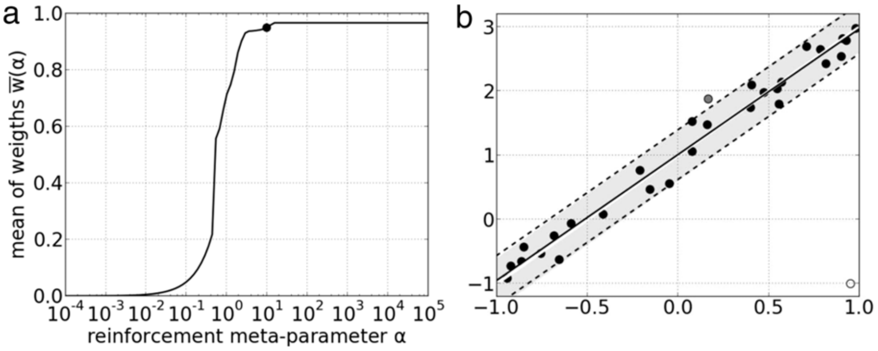


Fig. 11. Reinforcement meta-parameter selection for linear regression with L_1 regularisation on PPRs. The left panel shows the mean of weights in terms of α for optimal model parameters. The right panel shows the linear regression which is obtained with $\alpha = 10.19$ (black line), with respect to the true function (white line). The 30 data are shown by circles whose darkness is proportional to their respective weights. The estimated and true 95% confidence intervals are shown by dashed lines and the shaded region, respectively.

and are less sensitive to e.g. outliers. This important problem is discussed in this section.

5.1. Meta-parameter optimisation schemes

Two common approaches to deal with meta-parameters are validation and Bayesian priors. Validation consists in choosing the best meta-parameter with respect to the performances obtained by the model on test data. This approach includes cross-validation, leave-one-out validation, bootstrap, etc. However, it is in practice impossible to obtain test data which are guaranteed to be clean from outliers. Hence, since using PPRs produces models which are less sensitive to AFDs and may therefore give them very small probabilities, a validation approach would probably choose a very large α value, what is undesirable. It is not easy either to define a sensible Bayesian prior for the α parameter. This paper proposes an alternative way to optimise α. As discussed in Section 3.5, the instance weights measure the contribution of each observation to the parameter estimation and their mean can be seen as the percentage of the training sample which is actually used. Equivalently, the mean of the abnormality degrees a_i can be seen as the percentage of observations which are considered as AFDs. Hence, a simple solution to the α optimisation problem consists in using the α value which correspond to a plausible percentage of data supporting the model and a plausible percentage of outliers. A good choice is for example 95% for the former quantity or, equivalently, 5% for the latter. Several works suggest that it is more harmful to keep too many outliers than to remove too many correct data (Brodley & Friedl, 1999). Moreover, in the literature, real-world databases are estimated to contain around 5% of encoding errors (Maletic & Marcus, 2000; Redman, 1998). Of course, if prior knowledge suggests that more or less outliers are present in data, another percentage could be used.

Along this idea, meta-parameter α can be adapted as follows. Model parameters are initialised using a maximum likelihood approach, what corresponds to an infinite α value. Then, in the first iteration, an α value is searched (see Sections 5.2 and 5.3) in order to produce PPRs which are consistent with the constraint

$$\overline{w}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} w_i \approx 0.95. \tag{71}$$

The resulting PPRs are used to optimise model parameters and the algorithm iterates until convergence. At each iteration, a new α value is computed, since the weights w_i have changed meanwhile. In the end, PPRs and model parameters are obtained, whereas the condition (71) is satisfied.

5.2. The L_1 regularised case

When L_1 regularisation is used, observations must be first ordered according to their probability $p(x_i|\boldsymbol{\theta})$. Indeed, it follows from (13) and (34) that any observation whose probability is smaller than $\frac{1}{\alpha}$ has a weight $w_i = \alpha p(x_i|\boldsymbol{\theta})$, whereas all other instances have unitary weight. Then, the α search consists in looking for the smallest number k of observations with subunitary weight such that

$$\begin{aligned}\overline{w}(\alpha) &= \frac{1}{n}\sum_{i=1}^{n} w_i \\ &= \frac{1}{n}\sum_{i=1}^{k} \frac{1}{p(x_{k+1}|\boldsymbol{\theta})} p(x_i|\boldsymbol{\theta}) + \frac{1}{n}\sum_{i=k+1}^{n} 1 \leq 0.95, \end{aligned} \tag{72}$$

where α has been replaced by $1/p(x_{k+1}|\boldsymbol{\theta})$ since x_{k+1} is one of the instances with unitary weight $w_{k+1} = 1 = \alpha p(x_{k+1}|\boldsymbol{\theta})$. Since the $n-k$ last observations (and only them) necessarily have a unitary weight for $\overline{w}(\alpha) = 0.95$, the value of α which satisfies (71) can be estimated as

$$\alpha \approx \frac{\frac{k}{n} - 0.05}{\frac{1}{n}\sum_{i=1}^{k} p(x_i|\boldsymbol{\theta})}. \tag{73}$$

Fig. 11 shows an example of reinforcement meta-parameter choice for linear regression. Fig. 11(a) shows the mean of weights in terms of α for optimal model parameters, whereas Fig. 11(b) shows the linear regression which is obtained for the mean of weights $\overline{w}(\alpha) = 0.95$ with $\alpha = 10.19$. This solution is obtained with 3 iterations of the two-step iterative algorithm proposed in Sections 3 and 4.1.

5.3. The L_2 regularised case

For L_2 regularisation, a dichotomy approach can be used. Indeed, the mean of the weights is an increasing function of α, since its first-order derivative can easily be shown to be always positive. Initially, two very small and very large initial values of α (e.g. $\alpha_1 = 10^{-4}$ and $\alpha_3 = 10^{20}$) are picked. Then, one computes the intermediate value $\alpha_2 = \sqrt{\alpha_1\alpha_3}$ and only the two values $\alpha_i < \alpha_j$ such that $0.95 \in \left[\overline{w}(\alpha_i), \overline{w}(\alpha_j)\right]$ are kept. The algorithm is iterated until α_1 and α_3 are close enough, where the value α_2 can be chosen, since (71) is sufficiently close to be satisfied for $\overline{w}(\alpha_2)$. Fig. 12(a) shows the mean of weights in terms of α for optimal model parameters, whereas Fig. 12(b) shows the linear regression which is obtained for the mean of weights $\overline{w}(\alpha) = 0.95$ with $\alpha = 323.08$. This solution is obtained with 18 iterations of the two-step iterative algorithm proposed in Sections 3 and 4.1.

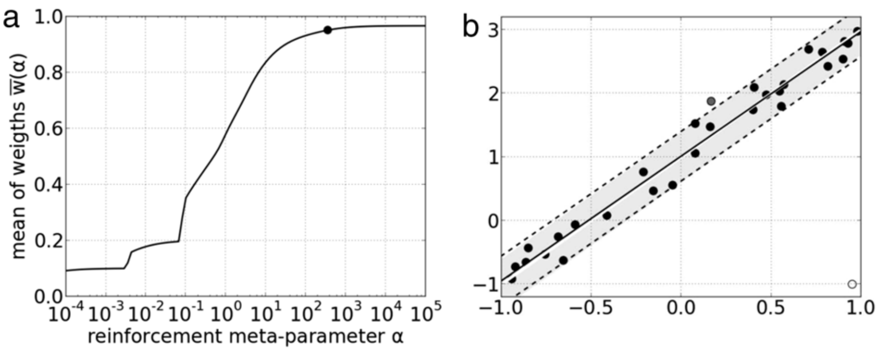


Fig. 12. Reinforcement meta-parameter selection for linear regression with L_2 regularisation on PPRs. The left panel shows the mean of weights in terms of α for optimal model parameters. The right panel shows the linear regression which is obtained with $\alpha = 323.08$ (black line), with respect to the true function (white line). The 30 data are shown by circles whose darkness is proportional to the sample weights. The estimated and true 95% confidence intervals are shown by dashed lines and the shaded region, respectively.

5.4. *Computational cost of reinforcing probabilities*

This section analyses the computational cost of the PPR methodology, with the above method for the choice of the reinforcement meta-parameter α. At each iteration of the algorithm proposed in Section 3.2, irrespective of the model type, three steps are performed.

First, the probabilities $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ are computed using the current estimate $\boldsymbol{\theta}^{\text{old}}$ of the model parameters. The computational cost depends on the considered model and is therefore difficult to characterise precisely. However, evaluating the probabilities $p\left(x_i|\boldsymbol{\theta}^{\text{old}}\right)$ is expected to be much faster than learning the estimate $\boldsymbol{\theta}^{\text{old}}$ itself. For example, in the case of linear regression, the former only requires to estimate the model output for each observation and to compute its probability, whereas the latter requires to compute a pseudo-inverse.

Second, the reinforcement meta-parameter α is optimised and the PPRs r_i are obtained from the probabilities. The computational cost depends on the type of regularisation. For L_1 and L_2 regularisation, Eqs. (12) and (17) provide cheap closed-form expression for the PPRs. With L_1 regularisation, the reinforcement meta-parameter is first estimated as explained in Section 5.2 and then the PPRs are computed. With L_2 regularisation, the PPRs must be computed at each step of the reinforcement meta-parameter search. The computational cost with L_2 regularisation is higher, but the computation of the reinforcement meta-parameter and of the PPRs only consists of simple operations.

Third, the instance weights w_i are computed and the model parameters are optimised with a weighted log-likelihood algorithm. Whereas weights are obtained using the simple closed-form expression (34), the model parameter optimisation is the most computationally expensive step in the proposed methodology. Indeed, learning algorithms usually involve costly operations like matrix inversion, gradient descent or convex optimisation with often non-linear time complexities.

Overall, the cost of reinforcing probabilities is dominated by the optimisation of the model parameters, since all other operations involve computations which are comparatively more simple. In practice, only a few iterations of the proposed algorithm are necessary before convergence. For example, in the two problems discussed in Sections 5.2 and 5.3, 3 and 18 iterations are necessary to converge when L_1 or L_2 regularisation is used, respectively. Experimentally, it is observed that the number of iterations decreases as the number of training instances increases. For small sample sizes, modern machine learning techniques are fast enough to cope with training the model a few times. For iterative learning procedures like gradient descent or convex optimisation, the proposed methodology can be considerably sped up by using the model parameters $\boldsymbol{\theta}^{\text{old}}$ obtained at a given iteration as a seed for the model parameters optimisation in the next iteration. Convergence of such methods should be much faster this way.

6. Experiments for kernel ridge regression

The goal of this section is to assess the PPR methodology in a more comprehensive way than the simple artificial problems used in Section 4. Indeed, since the aim of this paper is to propose a generic approach to robust maximum likelihood inference, it is important to check whether reinforced inference is able to perform at least as well as existing robust methods. The particular case of the kernel ridge regression discussed in Section 4.2 is considered. The choice of the reinforcement meta-parameter α is performed as proposed in Section 5. This section shows that the resulting methodology is competitive with existing methods to detect outliers in real settings for kernel ridge regression.

6.1. *Experimental settings*

As shown in Section 4.2, similarly to linear regression, kernel ridge regression is sensitive to outliers (Liu et al., 2011; Suykens, De Brabanter et al., 2002; Wen et al., 2010). Hence, Suykens, De Brabanter et al. (2002) propose to use a weighted kernel ridge regression to reduce the influence of outliers. This method consists in performing a standard kernel ridge regression, computing the error variables $e_i = \alpha_i/\gamma$ and training a weighted kernel ridge regression with the instance weights

$$w_i = \begin{cases} 1 & \text{if } |e_i/\hat{s}| \leq c_1 \\ \dfrac{c_2 - |e_i/\hat{s}|}{c_2 - c_1} & \text{if } c_1 \leq |e_i/\hat{s}| \leq c_2 \\ 10^{-4} & \text{if } c_2 < |e_i/\hat{s}| \end{cases} \tag{74}$$

where c_1 and c_2 are two constants. As discussed in Section 4.2, $\alpha_j > 0$ are dual weights and $\gamma = \sigma_\beta^2/\sigma_\epsilon^2$ is a meta-parameter which controls the compromise between errors and model complexity. The robust estimate $\hat{s}$ of the standard deviation of the error variables e_i is

$$\hat{s} = \frac{IQR}{2 \times 0.6745} \tag{75}$$

where IQR is the difference between the 25% and the 75% percentiles. Here, the constants $c_1 = 2.5$ and $c_2 = 3$ are used as suggested in Suykens, De Brabanter et al. (2002).

This section compares the reinforced kernel ridge regression proposed in Section 4.2 to both standard and weighted kernel ridge regression. For each method, the RBF kernel $k(x, z) = \exp$

Table 1
List of datasets used for experiments, ordered by size.

	Size	Dimensionality	Source
Pollution	60	15	StatLib[7]
Pyrim	74	27	LIBSVM[4]
Wine	124	256	MLG[5]
Nelson	128	2	NIST[6]
Nitrogen	141	166	ASRG[1]
Enso	168	1	NIST[6]
Hardware	209	6	UCI[8]
Chwirut1	214	1	NIST[6]
Tecator	240	100	StatLib[7]
Gauss3	250	1	NIST[6]
Bodyfat	252	14	LIBSVM[4]
Yacht	308	6	UCI[8]
Auto-MPG	392	7	LIBSVM[4]
NO2	500	7	StatLib[7]
Housing	506	13	LIBSVM[4]
Cooling	768	8	UCI[8]
Heating	768	8	UCI[8]
Stock	950	9	LIAAD[3]
Concrete	1030	8	UCI[8]
Mortgage	1049	15	FRB[2]
MG	1385	6	LIBSVM[4]
Space-GA	3107	6	LIBSVM[4]

$\left(\|x - z\|^2/\sigma^2\right)$ is used. The regularisation meta-parameter γ and the kernel precision parameter σ^{-2} are selected using 10-fold cross-validation on the mean squared error (MSE). The possible values belong to a 20 × 20 logarithmic grid such that $\gamma \in \left[10^{-3}, 10^{6}\right]$ and $\sigma^{-2} \in \left[10^{-4}, 10^{2}\right]$. Each experiment is repeated 100 times in order to obtain statistically significant results. For each repetition, the dataset is randomly split into a 70% training set and a 30% test set. Three distinct experiments are discussed here, where training data are polluted by 2%, 5% and 10% of outliers, respectively. Outliers are created by replacing the target value for some of the training instances by random values uniformly drawn in $[-10, 10]$. Input and output variables are normalised so that each variable has the same impact during learning. Table 1 shows the details for the datasets, which are chosen to span different small training set sizes and come from the Analytical Spectroscopy Research Group of the University of Kentucky,[1] the Federal Reserve Bank of Saint-Louis website,[2] the Laboratory of Artificial Intelligence and Decision Support of the University of Porto,[3] the LIBSVM data repository,[4] the Machine Learning Group of the Université catholique de Louvain website,[5] the NIST statistical reference datasets,[6] the StatLib dataset archive[7] and the UCI machine learning repository[8] (Asuncion & N, 2007). Reinforced kernel ridge regression is allowed a maximum of 20 iterations to learn a robust model. The method proposed in Section 5 is used to choose the reinforcement meta-parameter under the hypothesis that data contain 5% of outliers, i.e. the value of α is tuned so that the mean of instance weights is $\overline{w}(\alpha) \approx 0.95$.

For each trained model, the generalisation error is estimated by the MSE over the test set. The results for the 100 repetitions are averaged and the Wilcoxon rank-sum (Riffenburgh, 2012) statistic is used to assess whether the different MSE distributions are similar or not. Small p-values mean that those distributions are significantly different; in this analysis, the significance threshold is 0.05. This test is preferred to e.g. a two sample test of means because the MSE values appear to have non-Gaussian distributions (Riffenburgh, 2012).

6.2. Experimental results

Tables 2–4 show the results obtained with 2%, 5% and 10% of outliers, respectively. In each table, the three first columns show the average MSE obtained using (i) standard kernel ridge regression on clean data, (ii) standard kernel ridge regression on polluted data and (iii) weighted kernel ridge regression on polluted data. Results show that, for each percentage of outliers, the performances of standard kernel ridge regression are altered by the outliers. Moreover, the weighted kernel ridge regression proposed in Suykens, De Brabanter et al. (2002) allows improving the results in the presence of outliers. These conclusions are supported by Wilcoxon rank-sum tests whose p-values are not given here, since they are all close to zero (i.e. the MSE distributions with the three methods are significantly different).

The fourth and sixth columns of Tables 2–4 show the results obtained with L_1 and L_2 regularised PPRs. The fifth (seventh) column shows the p-values for a Wilcoxon rank-sum test between the MSE distributions of the weighted kernel ridge regression and the L_1 (L_2) reinforced kernel ridge regression. With only 2% of outliers, the reinforced results are most of the time not significantly different than those obtained with weighted kernel ridge regression. In only a few cases, the reinforced results are slightly (and significantly) worse. With 5% of outliers, reinforced kernel ridge regression is always at least as good as weighted kernel ridge regression. In a few cases, the former (significantly) outperforms the latter. With 10% of outliers, our method is always significantly worse than Suykens' method. This is due to the hypothesis made by the method used to choose the reinforcement meta-parameter, i.e. that there are about 5% of outliers.

The p-values for the Wilcoxon rank-sum tests between the MSE distributions with L_1 and L_2 regularisation are shown in the last column of the tables. These large values show that there is no statistically significant difference between the results obtained by both regularisation schemes for each of the outlier percentages.

6.3. Conclusion of experiments

The experiments show that the PPR methodology works in realistic settings for kernel ridge regression. The method competes with Suykens, De Brabanter et al. (2002). In addition, whereas the weighted kernel ridge regression proposed in Suykens, De Brabanter et al. (2002) is an ad-hoc method which can only be used for kernelised regression methods, the PPR methodology can be used to robustify any maximum likelihood inference technique as shown in Section 4. Reinforced inference could therefore e.g. be used to deal with outliers in settings where it is less simple to design an ad-hoc method for robust inference.

The method described in Section 5 needs to make the hypothesis that data contain a known percentage x of outliers. Three situations can be distinguished in this context. First, when the actual percentage of outliers is smaller than x%, the results of reinforced inference remain often similar to those of robust inference methods. Second, when the actual percentage of outliers is close to x%, good results are obtained and reinforced inference may outperform robust inference. Third, when the actual percentage of outliers is larger than x%, reinforced inference does not remove enough outliers and is outperformed by robust inference. In conclusion, the experimental results confirm the discussion in Section 5.1, i.e. that it is more harmful to keep too many outliers than to remove too many correct data (Brodley & Friedl, 1999).

[1] http://kerouac.pharm.uky.edu/asrg/cnirs.
[2] http://www.stls.frb.org/fred/data/zip.html.
[3] http://www.dcc.fc.up.pt/ltorgo/Regression/DataSets.html.
[4] http://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/.
[5] http://www.ucl.ac.be/mlg/.
[6] http://www.itl.nist.gov/div898/strd/nls/nls_main.shtml.
[7] http://lib.stat.cmu.edu/datasets/.
[8] http://archive.ics.uci.edu/ml/index.html.

Table 2
Results obtained with standard, weighted and reinforced kernel ridge regression with 2% of outliers in training data. Average MSE for 100 repetitions and p-values for Wilcoxon rank-sum tests between the MSE distributions are shown. MSE distributions with L_1 and L_2 regularised PPRs are compared to those of weighted regression. MSEs in grey for regularised PPRs are significantly worse than MSEs with weighted kernel ridge regression. Performances of standard kernel ridge regression on clean data are shown as reference. See text for more details.

	MSE with clean data	MSE with polluted data	MSE with weighted	L_1 reinforced PPRs		L_2 reinforced PPRs		L_1 vs. L_2
				MSE	p-value	MSE	p-value	p-value
Pollution	4.6e−1	7.5e−1	5.7e−1	5.6e−1	0.88	5.6e−1	0.94	0.93
Pyrim	6.6e−1	9.5e−1	7.7e−1	7.2e−1	0.59	7.3e−1	0.75	0.86
Wine	9.0e−3	1.2e−1	3.5e−2	4.0e−2	0.40	4.3e−2	0.44	0.97
Nelson	1.2e−1	2.5e−1	2.3e−1	1.6e−1	0.66	1.7e−1	0.68	0.99
Nitrogen	2.0e−1	3.9e−1	4.4e−1	5.1e−1	0.54	4.4e−1	0.79	0.69
Enso	5.4e−1	7.4e−1	5.8e−1	6.4e−1	0.04	6.4e−1	0.08	0.64
Hardware	2.2e−1	2.5e−1	2.2e−1	2.2e−1	0.82	2.1e−1	0.90	0.97
Chriwut1	2.1e−2	4.6e−2	2.8e−2	2.8e−2	0.75	2.8e−2	0.75	0.99
Tecator	6.7e−2	2.9e−1	1.5e−1	9.7e−2	0.16	1.2e−1	0.21	0.79
Gauss3	3.6e−3	5.9e−2	2.0e−2	1.6e−2	0.79	1.4e−2	0.92	0.82
Bodyfat	2.7e−2	8.0e−2	2.7e−2	4.7e−2	0.97	2.7e−2	0.78	0.75
Yacht	2.9e−3	1.8e−1	2.5e−2	3.8e−2	0.00	3.4e−2	0.00	0.77
Auto-MPG	1.2e−1	1.6e−1	1.3e−1	1.3e−1	0.51	1.3e−1	0.70	0.84
NO2	4.5e−1	4.9e−1	4.7e−1	4.8e−1	0.13	4.8e−1	0.24	0.69
Housing	1.3e−1	2.2e−1	1.6e−1	1.8e−1	0.07	1.9e−1	0.03	0.68
Cooling	2.4e−2	1.1e−1	3.2e−2	3.3e−2	0.32	3.3e−2	0.07	0.41
Heating	2.3e−3	7.6e−2	5.6e−3	5.6e−3	0.46	5.6e−3	0.28	0.76
Stock	1.1e−2	5.4e−2	1.7e−2	1.6e−2	0.64	1.6e−2	0.84	0.69
Concrete	1.2e−1	2.1e−1	1.3e−1	1.3e−1	0.09	1.3e−1	0.06	0.81
Mortgage	4.9e−4	1.1e−2	2.7e−3	3.5e−3	0.90	2.8e−3	0.50	0.49
MG	2.8e−1	3.2e−1	2.9e−1	3.0e−1	0.01	3.0e−1	0.05	0.42
Spage-GA	2.8e−1	3.2e−1	2.9e−1	2.9e−1	0.34	2.9e−1	0.29	0.98

Table 3
Results obtained with standard, weighted and reinforced kernel ridge regression with 5% of outliers in training data. Average MSE for 100 repetitions and p-values for Wilcoxon rank-sum tests between the MSE distributions are shown. MSE distributions with L_1 and L_2 regularised PPRs are compared to those of weighted regression. MSEs in bold for regularised PPRs are significantly better than MSEs with weighted kernel ridge regression. Performances of standard kernel ridge regression on clean data are shown as reference. See text for more details.

	MSE with clean data	MSE with polluted data	MSE with weighted	L_1 reinforced PPRs		L_2 reinforced PPRs		L_1 vs. L_2
				MSE	p-value	MSE	p-value	p-value
Pollution	4.9e−1	7.9e−1	7.2e−1	7.5e−1	0.72	8.8e−1	0.55	0.80
Pyrim	5.6e−1	9.9e−1	6.2e−1	7.2e−1	0.55	7.2e−1	0.52	0.98
Wine	7.2e−3	2.1e−1	9.8e−2	7.4e−2	0.70	1.1e−1	0.76	0.92
Nelson	1.2e−1	4.5e−1	3.5e−1	3.7e−1	0.13	1.9e−1	0.17	0.96
Nitrogen	2.0e−1	4.6e−1	2.7e−1	2.7e−1	0.96	2.7e−1	0.78	0.86
Enso	5.4e−1	9.7e−1	6.3e−1	7.5e−1	0.98	6.2e−1	0.61	0.61
Hardware	2.4e−1	3.7e−1	2.9e−1	2.8e−1	0.60	3.0e−1	0.49	0.90
Chriwut1	2.1e−2	8.6e−2	5.3e−2	5.2e−2	0.58	5.1e−2	0.55	0.95
Tecator	6.0e−2	3.4e−1	1.6e−1	1.5e−1	0.69	1.5e−1	0.42	0.71
Gauss3	3.5e−3	1.1e−1	2.1e−2	2.1e−2	0.63	2.1e−2	0.86	0.47
Bodyfat	2.9e−2	1.7e−1	4.9e−2	5.6e−2	0.59	5.1e−2	0.96	0.50
Yacht	3.3e−3	2.6e−1	4.1e−2	**2.3e−2**	0.01	**2.4e−2**	0.03	0.67
Auto-MPG	1.3e−1	2.0e−1	1.4e−1	1.4e−1	0.30	1.4e−1	0.23	0.90
NO2	4.7e−1	5.3e−1	4.9e−1	4.9e−1	0.95	4.8e−1	0.87	0.81
Housing	1.4e−1	2.7e−1	1.8e−1	1.7e−1	0.86	1.8e−1	0.85	0.72
Cooling	2.4e−2	1.5e−1	4.2e−2	**3.8e−2**	0.00	**3.7e−2**	0.00	0.98
Heating	2.2e−3	1.1e−1	1.3e−2	**1.1e−2**	0.00	**1.0e−2**	0.00	0.94
Stock	1.1e−2	8.7e−2	2.3e−2	**2.1e−2**	0.01	**2.2e−2**	0.01	0.93
Concrete	1.2e−1	2.7e−1	1.5e−1	**1.4e−1**	0.02	**1.4e−1**	0.01	0.78
Mortgage	5.5e−4	2.4e−2	8.4e−3	7.1e−3	0.81	7.6e−3	0.82	0.98
MG	2.8e−1	3.4e−1	2.9e−1	2.9e−1	0.80	2.9e−1	0.80	0.99
Spage-GA	2.8e−1	3.4e−1	3.1e−1	3.0e−1	0.22	3.0e−1	0.31	0.84

7. Conclusion

This paper introduces a generic method to deal with outliers in the case of probabilistic models. Indeed, it is well-known that maximum likelihood inference of model parameters is sensitive to abnormally frequent data. An approach is proposed to robustify maximum likelihood techniques. Probabilities are reinforced by pointwise probability reinforcements (PPRs), whose properties can be controlled by regularisation to obtain a compromise between fitting data and finding outliers. It is shown that L_1 regularisation induces sparse PPRs, what results in a few observations being considered as potential abnormally frequent data. Using L_2 regularisation, a similar, yet smoother solution is obtained. In addition, it is proven that observations which are more probable with respect to the parametric model are going to be less reinforced and correspond to larger reinforced probabilities. In order to perform maximum likelihood inference with PPRs, a generic two-step iterative algorithm is proposed which alternatively optimises the PPRs and the model parameters. For the PPRs, cheap closed-form update rules are obtained. For the model parameters, a lower bound to the objective function is derived which allows us to obtain an approximate maximisation step. This step consists

Table 4
Results obtained with standard, weighted and reinforced kernel ridge regression with 10% of outliers in training data. Average MSE for 100 repetitions and p-values for Wilcoxon rank-sum tests between the MSE distributions are shown. MSE distributions with L_1 and L_2 regularised PPRs are compared to those of weighted regression. MSEs in grey for regularised PPRs are significantly worse than MSEs with weighted kernel ridge regression. Performances of standard kernel ridge regression on clean data are shown as reference. See text for more details.

	MSE with clean data	MSE with polluted data	MSE with weighted	L_1 reinforced PPRs		L_2 reinforced PPRs		L_1 vs. L_2
				MSE	p-value	MSE	p-value	p-value
Pollution	4.7e−1	9.4e−1	6.4e−1	8.1e−1	0.00	7.8e−1	0.01	0.86
Pyrim	5.1e−1	9.4e−1	7.0e−1	8.3e−1	0.02	8.4e−1	0.02	0.95
Wine	9.6e−3	3.9e−1	1.9e−1	2.3e−1	0.00	2.4e−1	0.00	0.89
Nelson	4.4e−1	4.8e−1	3.7e−1	4.2e−1	0.02	3.8e−1	0.02	0.89
Nitrogen	1.9e−1	5.6e−1	3.2e−1	4.1e−1	0.00	4.1e−1	0.00	0.88
Enso	5.4e−1	1.2	1.0	1.0	0.01	1.0	0.02	0.82
Hardware	2.0e−1	3.8e−1	2.7e−1	2.5e−1	1.00	2.5e−1	0.75	0.68
Chriwut1	2.1e−2	1.5e−1	6.3e−2	8.3e−2	0.00	8.4e−2	0.00	0.89
Tecator	6.7e−2	4.5e−1	1.7e−1	2.8e−1	0.00	2.5e−1	0.00	0.95
Gauss3	3.7e−3	2.2e−1	4.8e−2	9.0e−2	0.00	9.0e−2	0.00	0.47
Bodyfat	3.0e−2	2.7e−1	7.1e−2	1.3e−1	0.00	1.3e−1	0.00	0.89
Yacht	2.9e−3	4.4e−1	1.4e−1	1.8e−1	0.00	1.8e−1	0.00	0.68
Auto-MPG	1.3e−1	2.5e−1	1.7e−1	1.9e−1	0.00	1.8e−1	0.00	0.85
NO2	4.7e−1	5.9e−1	4.9e−1	5.1e−1	0.19	5.0e−1	0.26	0.88
Housing	1.3e−1	3.8e−1	2.0e−1	2.4e−1	0.00	2.4e−1	0.00	0.80
Cooling	2.4e−2	1.8e−1	7.1e−2	1.2e−1	0.00	1.2e−1	0.00	0.88
Heating	2.3e−3	1.5e−1	4.4e−2	9.1e−2	0.00	8.9e−2	0.00	0.65
Stock	1.1e−2	1.5e−1	3.7e−2	7.0e−2	0.00	7.0e−2	0.00	0.99
Concrete	1.2e−1	3.3e−1	1.8e−1	2.3e−1	0.00	2.3e−1	0.00	0.79
Mortgage	5.5e−4	5.1e−2	3.3e−2	2.9e−2	0.47	2.9e−2	0.49	0.96
MG	2.8e−1	3.8e−1	3.0e−1	3.2e−1	0.00	3.2e−1	0.00	0.63
Spage-GA	2.9e−1	3.7e−1	3.3e−1	3.4e−1	0.01	3.4e−1	0.02	0.82

in maximising a weighted log-likelihood. The instance weights are obtained from the PPRs and can also be used by experts to analyse data in search of outliers. The adaptation of four standard probabilistic techniques (linear regression, kernel ridge regression, logistic regression and PCA) shows the generality of the proposed approach. Outliers can be detected in supervised or unsupervised contexts and a degree of abnormality can be obtained for each observation. The average degree of abnormality can be easily controlled using a meta-parameter α, what can be used to select an adequate compromise in the regularisation. Algorithms are proposed to choose a good intermediate value of α in the case of L_1 and L_2 regularisation.

Experiments for kernel ridge regression show that the methodology proposed in this paper works in realistic settings. The obtained results depend on the choice of the PPR regularisation meta-parameter α, i.e. on the estimated percentage of outliers if α is chosen so that the mean weight of instances during inference corresponds to that percentage. When the percentage of outliers is overestimated, reinforced inference provides good results which are comparable to those of existing methods. Also, when the percentage of outliers is accurately estimated, reinforced inference may outperform other methods. Eventually, when the percentage of outliers is underestimated, the performances of reinforced inference decrease. In conclusion, reinforced inference can be used in practice and meets the goal of this paper, i.e. to provide a generic method which can be used to robustify maximum likelihood inference techniques. Experiments show that it is necessary to have at least a rough upper bound on the percentage of outliers. Using PPRs, there is no need for ad-hoc procedures like e.g. those proposed in the literature for kernel ridge regression (Jiyan et al., 2011; Liu et al., 2011; Suykens, De Brabanter et al., 2002; Wen et al., 2010), logistic regression (Rousseeuw & Christmann, 2003) and principal component analysis (Fan et al., 2011; Huber, 1981).

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments which allowed them to significantly improve the paper. Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11.

References

Aitkin, M., & Wilson, G. T. (1980). Mixture models, outliers, and the EM algorithm. *Technometrics*, *22*(3), 325–331.

Archambeau, C., Delannay, N., & Verleysen, M. (2006). Robust probabilistic projections. In *Proceedings of the 23rd int. conf. on machine learning* (pp. 33–40).

Asuncion, A., & Newman, D.J. (2007). UCI machine learning repository. URL http://www.ics.uci.edu/~mlearn/MLRepository.html.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge, UK: Cambridge University Press.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. New York, NY: Wiley.

Beckman, R. J., & Cook, R. D. (1983). Outlier..........s. *Technometrics*, *25*(2), 119–149.

Bernardo, J., & Smith, A. (2007). *Bayesian theory*. New York, NY: Wiley.

Bishop, C. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Brodley, C. E., & Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, *11*, 131–167.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: a survey. *ACM Computing Surveys*, *41*(3), 15:1–15:58.

Chen, D., & Jain, R. (1994). A robust backpropagation learning algorithm for function approximation. *IEEE Transactions on Neural Networks*, *5*(3), 467–479.

Cheng, T.-C., & Biswas, A. (2008). Maximum trimmed likelihood estimator for multivariate mixed continuous and categorical data. *Computational Statistics & Data Analysis*, *52*(4), 2042–2065.

Chuang, C.-C., Su, S.-F., & Hsiao, C.-C. (2000). The annealing robust backpropagation (arbp) learning algorithm. *IEEE Transactions on Neural Networks*, *11*(5), 1067–1077.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, *74*(365), 169–174.

DasGupta, A. (2011). *Probability for statistics and machine learning*. (pp. 498–521). Berlin: Springer, The exponential family and statistical applications (Chapter).

Daszykowski, M., Kaczmarek, K., Heyden, Y. V., & Walczak, B. (2007). Robust statistics in data analysis—a review: basic concepts. *Chemometrics and Intelligent Laboratory Systems*, *85*(2), 203–219.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *39*(1), 1–38.

Domingo, C., & Watanabe, O. (2000). Madaboost: a modification of adaboost. In *Proceedings of the 13th ann. conf. on computational learning theory* (pp. 180–189).

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York, NY: Wiley.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*, 407–499.

Eskin, E. (2000). Anomaly detection over noisy data using learned probability distributions. In *Proceedings of the 17th int. conf. on machine learning* (pp. 255–262).

Fan, Z., Liu, E., & Xu, B. (2011). Weighted principal component analysis. In *Proceedings of the 3rd int. conf. on artificial intelligence and computational intelligence, part III* (pp. 569–574).

Filzmoser, P., Maronna, R., & Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, *52*(3), 1694–1711.

Ganapathiraju, A., Picone, J., & State, M. (2000). Support vector machines for automatic data cleanup. In *Proceedings of the 6th int. conf. on spoken language processing* (pp. 210–213).

Guyon, I., Matic, N., & Vapnik, V. (1996). Discovering informative patterns and data cleaning. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining* (pp. 181–203). Cambridge, MA: AAAI, MIT Press.

Hadi, A. S., & Luceo, A. (1997). Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Computational Statistics & Data Analysis*, *25*(3), 251–272.

Hadi, A. S., & Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, *88*(424), 1264–1272.

Hawkins, D. M. (1980). *Identification of outliers*. London, UK: Chapman and Hall.

Hernandez-Lobato, D., Hernandez-Lobato, J.M., & Dupont, P. (2011). Robust multi-class gaussian process classification. In *Advances in neural information processing systems, Vol. 24* (pp. 280–288).

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hu, F., & Zidek, J. V. (2002). The weighted likelihood. *Canadian Journal of Statistics*, *30*(3), 347–371.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, *35*(1), 73–101.

Huber, P. J. (1981). *Robust statistics*. New York, NY: Wiley.

Hubert, M., Rousseeuw, P., & Verdonck, T. (2009). Robust PCA for skewed data and its outlier map. *Computational Statistics & Data Analysis*, *53*(6), 2264–2274.

Jiyan, H., Guan, G., & Qun, W. (2011). Robust location algorithm based on weighted least-squares support vector machine (WLS-SVM) for non-line-of-sight environments. *International Journal of the Physical Sciences*, *6*(25), 5897–5905.

Kowalczyk, A., Smola, A. J., & Williamson, R. C. (2001). Kernel machines and boolean functions. In *Advances in neural information processing systems, Vol. 14* (pp. 439–446). British Columbia, Canada: Vancouver.

Lawrence, N.D., & Schölkopf, B. (2001). Estimating a kernel fisher discriminant in the presence of label noise. In *Proceedings of the 18th int. conf. machine learning* (pp. 306–313).

Liano, K. (1996). Robust error measure for supervised neural network learning with outliers. *IEEE Transactions on Neural Networks*, *7*(1), 246–250.

Liu, J., Li, J., Xu, W., & Shi, Y. (2011). A weighted lq adaptive least squares support vector machine classifiers robust and sparse approximation. *Expert Systems with Applications*, *38*(3), 2253–2259.

Maletic, J.I., & Marcus, A. (2000). Data cleansing: beyond integrity analysis. In *Proceedings of the conf. on information quality* (pp. 200–209).

Muller, K. R., Mika, S., Ratsch, G., Tsuda, K., & Scholkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, *12*(2), 181–201.

Neykov, N., Filzmoser, P., Dimova, R., & Neytchev, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, *52*(1), 299–308.

Redman, T. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, *2*(2), 79–82.

Rekaya, R., Weigel, K. A., & Gianola, D. (2001). Threshold model for misclassified binary responses with applications to animal breeding. *Biometrics*, *57*(4), 1123–1129.

Riffenburgh, R. (2012). *Statistics in medicine*. Academic Press.

Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*(388), 871–880.

Rousseeuw, P., & Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis*, *43*(3), 315–332.

Rousseeuw, P., & Van Driessen, K. (2006). Computing lts regression for large data sets. *Data Mining and Knowledge Discovery*, *12*, 29–45.

Ruppert, D., & Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, *75*(372), 828–838.

Saunders, C., Gammerman, A., & Vovk, V. (1998). Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th int. conf. on machine learning* (pp. 515–521).

Simeckova, N. (2005). Maximum weighted likelihood estimator in logistic regression. In *Proceedings of the 14th ann. conf. of doctoral students* (pp. 144–148).

Stanimirova, I., Walczak, B., Massart, D., & Simeonov, V. (2004). A comparison between two robust PCA algorithms. *Chemometrics and Intelligent Laboratory Systems*, *71*(1), 83–95.

Suykens, J., De Brabanter, J., Lukas, J., & Vandewalle, J. (2002). Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, *48*, 85–105.

Suykens, J., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific.

Tipping, M. E., & Bishop, C. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*, 611–622.

Wen, W., Hao, Z., & Yang, X. (2010). Robust least squares support vector machine based on recursive outlier elimination. *Soft Computing*, *14*(11), 1241–1251.

Xu, L., & Yuille, A. (1995). Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Transactions on Neural Networks*, *6*(1), 131–143.

Zhang, W., Rekaya, R., & Bertrand, K. (2006). A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer. *Bioinformatics*, *22*(3), 317–325.

Zhu, X., & Wu, X. (2004). Class noise vs. attribute noise: a quantitative study. *Artificial Intelligence Review*, *22*, 177–210.