Abstract

Robustness to specification of the error distribution is an important concern in regression. In this paper, we propose a general nonparametric scale mixture model for the error distribution. For fitting such mixtures, the predictive recursion method is a simple and computationally efficient alternative to existing methods. Here, a predictive recursion-based likelihood function is constructed, and estimation of the regression parameters proceeds by maximizing this function. A hybrid predictive recursion–EM algorithm is proposed for this purpose. The method’s performance is compared with that of existing methods in simulations and real data analyses.

Keywords and phrases: EM algorithm; Dirichlet process; marginal likelihood; nonparametric maximum likelihood; normal scale mixture; profile likelihood.

1 Introduction

Consider the standard linear regression model,

$$y = X\beta + \varepsilon,$$

where $y = (y_1, \ldots, y_n)^T$ is a $n \times 1$ vector of response variables, $X$ is a $n \times p$ matrix of predictor variables, with $i$th row $x_i = (x_{i1}, \ldots, x_{ip})^T$, $\beta$ is a $p \times 1$ vector of regression coefficients, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an $n \times 1$ vector of iid errors with common density $f$. In classical linear model applications, one assumes that $f$ is a normal distribution with mean zero and unknown variance $\sigma^2$. In this case, the ordinary least squares method provides the optimal estimates of $(\beta, \sigma^2)$. However, if $f$ has heavier-than-normal tails,
then the accuracy of the least squares solutions is lost. This is due to the sensitivity of
the least squares method to residuals with large magnitude.

When the error distribution may be non-normal, one may consider a robust method
that is less sensitive to outliers. Standard approaches to robust regression fall under the
category of M-estimation (e.g., Huber 1973, 1981), and include methods based on mini-
mizing an objective function different from the sum of squared residuals, such as least
absolute deviation, or $L_1$ regression. Surveys of these standard techniques, including
outlier detection, are given in Roussieuw and Leroy (1987) and Ryan (2002). However,
a likelihood-based method might be preferred, and a common approach is to model $f$
as a Student-t distribution, with small degrees of freedom, with an additional unknown
scale parameter. See, for example, Lange et al. (1989), Liu (1996), and Pinheiro et al.
(2001). The standard implementation uses the expectation–maximization (EM) algo-
rithm (Dempster et al. 1977), and the key is a representation of the Student-t distribution
as a scale mixture of normals (Andrews and Mallows 1974; West 1987).

Motivated by the scale mixture representation in the Student-t case, we consider a
more general model specified by an arbitrary mixing distribution $\Psi$, supported on $(0, \infty)$,
for the normal scale parameter. That is, we write the error density $f$ as a mixture
\begin{equation}
    f(\varepsilon) = \int_0^\infty N(\varepsilon \mid 0, u^2) \Psi(du). \tag{1}
\end{equation}
The density $f$ is symmetric about zero and has heavier than normal tails. For example,
if $\Psi$ is an inverse chi-square distribution, then $f$ is a Student-t density. Since the mixing
distribution $\Psi$ is completely unspecified, an additional scale parameter would not be
identifiable, so we take $(\beta, \Psi)$ as the only unknown parameters in our model.

To fit this semiparametric model, estimation of both $\beta$ and $\Psi$ is required. One
approach is maximum likelihood. More precisely, if $L(\beta, \Psi)$ is the joint likelihood for
$(\beta, \Psi)$, we define a profile likelihood function as $L_p(\beta) = L(\beta, \hat{\Psi}_\beta)$, where $\hat{\Psi}_\beta$
is the conditional maximum likelihood estimator of $\Psi$ for the given $\beta$. Then $L_p(\beta)$ can be
treated like a usual likelihood function, to produce estimators, tests, or confidence regions
for $\beta$. Existing algorithms for nonparametric maximum likelihood estimation of mixing
distributions (e.g., Wang 2007) can be used to compute $\hat{\Psi}_\beta$ and, in turn, the profile
likelihood $L_p(\beta)$. However, this profile likelihood generally has many local modes, so
optimization is challenging.

For a Bayesian approach, one could construct a marginal likelihood for $\beta$ by first
introducing a prior distribution for $\Psi$. A reasonable choice would be to take a Dirich-
let process prior for $\Psi$ (e.g. Ferguson 1973; Lo 1984; Müller and Quintana 2004). The idea is to integrate out $\Psi$ from the joint likelihood $L(\beta, \Psi)$ with respect to the prior,
leaving a marginal likelihood function $L_m(\beta)$ for $\beta$. Markov chain Monte Carlo algo-
rithms are available for evaluating this marginal likelihood (e.g., Escobar and West 1995;
MacEachern and Müller 1998; Neal 2000; Carvalho et al. 2010) but this is expensive be-
cause, in general, each marginal likelihood evaluation requires its own Monte Carlo run,
and optimization requires several such runs.

For a computationally efficient alternative, we consider an extension of the pre-
dictive recursion (PR) method, as discussed in Newton et al. (1998), Newton (2002),
Ghosh and Tokdar (2006), Martin and Ghosh (2008), Tokdar et al. (2009), and Martin and Tokdar
(2009). The PR algorithm is designed for fast nonparametric estimation of a mixing
distribution; see Section 2.1. Here, our model is a semiparametric mixture, where $\beta$ is the parameter of interest, and the mixing distribution $\Psi$ is a nuisance parameter. Martin and Tokdar (2011) proposed a general PR-driven methodology for inference in semiparametric mixtures. The idea is to construct an approximate marginal likelihood, $L_{pr}(\beta)$, defined in (4), to be used for inference on the interest parameter $\beta$; see Section 2.2. Details and the definition of $L_{pr}(\beta)$ in the regression context are presented in Section 3, along with a comparison with similar nonparametric likelihood and Bayesian approaches. Optimization of $L_{pr}(\beta)$ is discussed in Section 4. The recommended strategy is a hybrid PR–EM algorithm that takes advantage of the latent scale parameter representation of the mixture model (2). This hybrid algorithm is fast and easy to compute, and is shown in Section 5 to be quite accurate compared to existing methods. Moreover, like many other robust regression methods, PR–EM produces, as a by-product, a set of weights which can be used for outlier detection. Section 6 provides some concluding remarks.

2 Predictive recursion

2.1 Nonparametric mixtures

PR is a fast, stochastic algorithm, designed for estimating mixing distributions in nonparametric mixture models. It was first proposed as an alternative to Markov chain Monte Carlo methods in fitting Bayesian Dirichlet process mixture models (Newton et al. 1998; Newton 2002). To summarize the general case, let $y_1, \ldots, y_n$ be iid with density $f(y)$, where $f = f_\psi$ is modeled as a mixture $\int_{\mathcal{U}} p(y \mid u) \psi(u) \lambda(du), p(y \mid u)$ is a known kernel, and $\psi$ is an unknown density with respect to a dominating $\sigma$-finite measure $\lambda$ on $\mathcal{U}$. The PR algorithm estimates $\psi$ and $f_\psi$ as follows.

**PR Algorithm.** Initialize the algorithm by choosing a $\lambda$-density $\psi_0$ and a sequence of weights $\{w_i : i \geq 1\} \subset (0, 1)$. For $i = 1, \ldots, n$, repeat the following steps.

1. Compute the mixture density:
   \[
   f_{i-1}(\cdot) \equiv f_{\psi_{i-1}}(\cdot) = \int p(\cdot \mid u) \psi_{i-1}(u) \lambda(du). \tag{2}
   \]

2. Update the mixing density estimate:
   \[
   \psi_i(u) = (1 - w_i) \psi_{i-1}(u) + w_i \frac{p(y_i \mid u) \psi_{i-1}(u)}{f_{i-1}(y_i)}. \tag{3}
   \]

Return $\psi_n$ and $f_n = f_{\psi_n}$ as the PR estimates of $\psi$ and $f_\psi$, respectively.

Two key properties of the PR algorithm are speed and ease of implementation. Also, PR is able to produce an estimate of the mixing distribution which has a density with respect to the prescribed dominating measure $\lambda$. Compare this to the nonparametric maximum likelihood estimate which is almost surely discrete (Lindsay 1995).

Large-sample convergence properties of the PR estimates are given in Tokdar et al. (2009). In particular, under suitable regularity conditions, the PR estimate $f_n$ of the
mixture density is consistent and, if the mixing distribution is identifiable, then the PR estimate \(\psi_n\) of the mixing density is also consistent. Bounds on the PR rate of convergence, in the well- and misspecified model cases, are presented in Martin and Tokdar (2009).

To end this subsection, we discuss a few specific properties of the PR algorithm that are relevant to its implementation.

- The weights \(\{w_i\}\) in the PR algorithm are required to satisfy \(\sum_{i=1}^{\infty} w_i = \infty\) and \(\sum_{i=1}^{\infty} w_i^2 < \infty\). Subject to these conditions, the practical performance of PR is not too sensitive to the particular choice. Here we take \(w_i = (i + 1)^{-1}\).

- The PR estimates depend on the order in which the data are processed. This dependence can be weakened by averaging the PR estimates over several (random) permutations of the data sequence. In our experience, averaging over 25 permutations is sufficient (Martin and Tokdar 2012) and, given the speed of PR, this does not significantly increase the computational cost.

### 2.2 Semiparametric mixtures

As an extension of the nonparametric mixture model setup in the previous subsection, consider the case where the kernel \(p(y \mid u)\) depends on an unknown parameter \(\theta\), i.e., \(p(y \mid u) = p(y \mid \theta, u)\). In this context, the structural parameter \(\theta\) is typically of primary interest, while the mixing density \(\psi\) is a nuisance parameter.

For this problem, Martin and Tokdar (2011) proposed an extension of the PR algorithm that produces a sort of likelihood function for \(\theta\) to be used for inference. Let \(f_{k,\theta}\) be the PR mixture density estimates based on data \((y_1, \ldots, y_k), k = 1, \ldots, n, \) and kernel \(p(y \mid \theta, u)\), where \(\theta\) is taken to be fixed. Consider the function

\[
L_{PR}(\theta) = \prod_{i=1}^{n} f_{i-1,\theta}(y_i).
\]

This function is called the PR marginal likelihood for \(\theta\). Despite its familiar product-of-densities form, this is not a genuine likelihood function for \(\theta\) under the posited semiparametric model. Martin and Tokdar (2011) use PR’s natural connection to the Bayesian Dirichlet process prior model to argue that \(L_{PR}(\theta)\) is an approximate marginal likelihood for \(\theta\). They also demonstrate the large-sample convergence properties of \(L_{PR}(\theta)\), and give some examples. Martin and Tokdar (2012) and Martin (2013) give other applications.

### 3 PR marginal likelihood for regression

Here we specialize the PR marginal likelihood methodology to the regression context. That is, consider the linear model \(y = X\beta + \varepsilon\), where \(\varepsilon\) is an \(n\)-vector of iid errors assumed to have density \(f\) of the mixture form in (1), where the mixing density \(\psi\), supported on \(\mathcal{U} \subseteq (0, \infty)\) is unknown. As discussed in Section 1, \(f\) has heavier-than-normal tails, so inference on \(\beta\) based on such a model will be less sensitive to extreme
observations compared to inference based on a basic normal model. To put this in the form suitable for PR, write the mixture model for the residuals,

\[ f(y_i - x_i^\top \beta) = \int_{U} N(y_i - x_i^\top \beta \mid 0, u^2) \psi(u) \, du, \quad i = 1, \ldots, n. \]

With this representation, it is clear that we can apply the PR algorithm to the \( \beta \)-dependent residuals, \( y_i - x_i^\top \beta \), to estimate the mixing and mixture densities. If \( f_{k, \beta} \), \( k = 1, \ldots, n \), are the PR estimates of the mixture density for the given \( \beta \), the following PR marginal likelihood for \( \beta \) obtains:

\[ L_{pr}(\beta) = \prod_{i=1}^{n} f_{i, \beta}^{-1}(y_i - x_i^\top \beta). \tag{4} \]

This is fast and easy to compute. As with all likelihood functions, we propose to estimate \( \beta \) by maximizing this PR marginal likelihood or, equivalently, the PR log-marginal likelihood \( \ell_{pr}(\beta) = \log L_{pr}(\beta) \). More on this in Section 4.

There are two important remarks to be made concerning implementation of the PR marginal likelihood approach.

- For PR computations, a compact support \( U \subset (0, \infty) \) is required. We take \( U = [U_{\text{min}}, U_{\text{max}}] \), where \( U_{\text{min}} \) is fixed at \( 10^{-5} \), and \( U_{\text{max}} \) is to be specified. Since \( U_{\text{max}} \) helps to determine the overall scale of the error distribution, we should select \( U_{\text{max}} \) to satisfy two criteria. First, \( U_{\text{max}} \) should be sufficiently large so that the support is not overly restricted. Second, if the errors are actually normal, then the PR method should be able to recover this by producing an estimate of \( \psi \) that is tightly concentrated around the usual root mean square error estimator \( \hat{\sigma} \) of \( \sigma \). A simple idea is to take \( U_{\text{max}} = \max \{50, 3\hat{\sigma} \} \).

- For the initial guess \( \psi_0 \) of the mixing density, there are many possibilities. Here we make a “non-informative” choice, taking \( \psi_0 \) to be a uniform density on \( U \). One could also consider an “informative” choice of \( \psi_0 \), e.g., a gamma density, truncated to \( U \), with mode at the least squares estimator \( \hat{\sigma} \) of the normal scale \( \sigma \).

- As discussed previously, to weaken the dependence of the PR estimates on the data ordering, we recommend averaging over 25 data permutations. These permutations can be selected at random, but it is important that the permutations remain fixed over the optimization process. If new permutations are used for each \( L_{pr}(\beta) \) evaluation, then the optimization process may not converge.

To end this section, we consider those two alternatives to the proposed PR-based methodology discussed in Section 1. The first is a profile likelihood approach based on nonparametric maximum likelihood. Mimicking the idea in the PR approach above, one can, for a fixed \( \beta \), calculate the residuals \( y - X \beta \), and apply some standard algorithm to get a conditional maximum likelihood estimate \( \hat{\Psi}_\beta \) for the mixing distribution, given \( \beta \), yielding a profile likelihood \( L_{p}(\beta) \) as in Section 1. Here we used a version of the fast algorithm in Wang (2007) for estimating the mixing distribution. We claim that this profile likelihood function is generally rough, so optimization over \( \beta \) is unstable and
computationally expensive. To justify this claim, we consider a simple special case of the regression problem with no predictor variables, i.e., iid data with location $\beta$. In this case, we can easily evaluate and plot the proposed marginal and profile likelihood functions. An independent sample of size $n = 100$ was drawn from a Student-t distribution with $df = 2$, centered at $\beta = 0$, and the corresponding likelihood functions for $\beta$ are plotted in Figure 1. The profile likelihood has a number of local modes, so numerical optimization is unstable. On the other hand, $L_{pr}(\beta)$ is smooth with one global mode, so numerical optimization is fast and easy.

A second alternative to PR is a genuine Bayesian approach with a Dirichlet process prior for $\Psi$. The naive approach would, for each fixed $\beta$, calculate the residuals $y - X\beta$ and then run a Markov chain Monte Carlo to integrate out the mixing distribution, yielding the marginal likelihood at $\beta$. To maximize this marginal likelihood over $\beta$, one requires a Monte Carlo simulation for each function evaluation. This is too expensive computationally to be of any practical use, despite the fact that efficient and automated software is available for inference in Dirichlet process mixture models (e.g., Jara et al. 2011). One can avoid repeatedly running Monte Carlo if $\beta$ is assigned a proper prior. That is, one can employ the technique in Chib (1995) to get a marginal likelihood for $\beta$ from a single, joint Monte Carlo run for $(\beta, \Psi)$. This single joint Monte Carlo is generally more expensive than the PR marginal likelihood strategy, so we do not explore this alternative here.

4 PR maximum likelihood regression

4.1 The estimator

To estimate $\beta$, we propose the PR maximum likelihood estimator $\hat{\beta}$, the maximizer of $L_{pr}(\beta)$ or $\ell_{pr}(\beta)$. For producing this estimator, one strategy is to use a prepackaged numerical optimization routine. We have found that, for relatively low-dimensional problems, this procedure works well, in terms of estimation accuracy and computational efficiency. However, for relatively high-dimensional problems, direct optimization seems
to be too costly, so we opt for a more efficient alternative; see Section 4.2.

Theoretical questions about existence and uniqueness of this maximum PR likelihood estimator are difficult to answer; this is a result of the complicated recursive structure of the PR algorithm. Martin and Tokdar (2011) make the conjecture that, under some conditions, \( \ell_{pr}(\beta) \) is a concave function of \( \beta \). Concavity would guarantee that a unique maximizer of \( L_{pr} \) could be found in practice. Moreover, concavity could also be used to establish asymptotic consistency of the PR maximum likelihood estimator (e.g., Hjort and Pollard 1993). A host of examples, including our Figure 1, support this conjecture, but no theory is currently available. See Section 6.

Robustness is a focus of this paper, so it is important to discuss in what sense the PR maximum likelihood estimator is robust. The key point is that the flexible scale mixture of normals model for the error terms allows for heavy tails so, if outlying observations are present, the model will be able to accommodate these and, therefore, the estimates based on this model will not be overly influenced by the outliers. It is in this sense that the PR estimator is robust. Furthermore, as we discuss in the upcoming subsection, maximizing the PR likelihood is equivalent to iteratively solving a weighted least squares problem. We demonstrate numerically in Section 5.2 that the weights assigned to observations are such that outliers tend to get weights near zero, further supporting the claim that extreme observations do not overly influence the PR estimates.

4.2 Computation: a hybrid PR–EM algorithm

The goal is to maximize the PR likelihood \( L_{pr}(\beta) \) or the log-likelihood \( \ell_{pr}(\beta) \). There is a computational gain that comes from taking advantage of the special structure of the problem. Along these lines, we present a hybrid PR–EM algorithm for maximizing \( L_{pr}(\beta) \). The jumping off point here is an alternative interpretation of the scale mixture formulation in (1) in terms of latent scale parameters \( U_1, \ldots, U_n \). Then we have the following trivial identity:

\[
\ell_{pr}(\beta) = \sum_{i=1}^{n} \log \int \mathcal{N}(y_i - x_i^T \beta \mid 0, u^2) \psi_{i-1,\beta}(u) \, du \\
= \sum_{i=1}^{n} \log \mathcal{N}(y_i - x_i^T \beta \mid 0, U_i^2) - \sum_{i=1}^{n} \log \left\{ \frac{\mathcal{N}(y_i - x_i^T \beta \mid 0, U_i^2)}{f_{i-1,\beta}(y_i - x_i^T \beta)} \right\}.
\]

Since this holds for all \( U_1, \ldots, U_n \), it must also hold if we take expectation with respect to some distribution over \( U_1, \ldots, U_n \). Our proposal is to integrate out \( U_i \) with respect to the density

\[
\psi^B_{i,\beta}(u) = \frac{\mathcal{N}(y_i - x_i^T \beta \mid 0, u^2) \psi_{i-1,\beta}(u)}{f_{i-1,\beta}(y_i - x_i^T \beta)},
\]
where \( \hat{\beta} \) is some estimate. This is exactly the Bayes posterior density based on “prior” \( \psi_{i-1, \hat{\beta}} \) and “data” \( y_i - x_i^\top \hat{\beta} \). In particular, write

\[
\ell_{pr}(\beta) = \sum_{i=1}^{n} \int \log \mathcal{N}(y_i - x_i^\top \beta \mid 0, u^2) \psi_{i, \beta}(u) \, du
\]

\[
- \frac{1}{2} \sum_{i=1}^{n} \int \log \left\{ \frac{\mathcal{N}(y_i - x_i^\top \beta \mid 0, u^2)}{f_{i-1, \beta}(y_i - x_i^\top \beta)} \right\} \psi_{i, \beta}(u) \, du.
\]

Write this as \( Q_1(\beta \mid \hat{\beta}) + Q_2(\beta \mid \hat{\beta}) \). Then \( Q_1(\beta \mid \hat{\beta}) \) simplifies to

\[
Q_1(\beta \mid \hat{\beta}) = -\frac{1}{2} \sum_{i=1}^{n} \hat{\omega}_i (y_i - x_i^\top \beta)^2 + \text{constant},
\]

where the weight \( \omega_i \), which depends on \( \psi_{i-1, \hat{\beta}} \) and \( y_i - x_i^\top \hat{\beta} \), is given by

\[
\hat{\omega}_i = \int u^{-2} \psi_{i, \beta}(u) \, du,
\]

(6)

the expected precision (inverse variance) under the distribution with density in (5). Both \( \hat{\omega}_i \) and the constant term depend on \( \hat{\beta} \), but not on \( \beta \).

We are now ready to state the proposed hybrid PR–EM algorithm. The key point is that maximizing \( Q_1(\beta \mid \hat{\beta}) \) corresponds to a weighted least squares problem, for which an analytic solution is available. The proposed algorithm produces a sequence of candidate estimators, \( \hat{\beta}^{(t)} \), \( t \geq 1 \), obtained by iteratively solving this weighted least squares problem, that generally increases \( \ell_{pr} \) at each iteration.

**PR–EM Algorithm.** Initialize the algorithm by choosing \( \hat{\beta} = \hat{\beta}^{(1)} \) and the input \( (\mathcal{W}, \psi_0, w_1, \ldots, w_n) \) for the PR portion. At iteration \( t \), repeat the following steps.

E. Compute the weights \( \hat{\omega}_1, \ldots, \hat{\omega}_n \) by running the PR algorithm with the residuals \( y_1 - x_1^\top \hat{\beta}^{(t)} \), \( \ldots \), \( y_n - x_n^\top \hat{\beta}^{(t)} \) as data.

M. Choose \( \hat{\beta}^{(t+1)} \) to maximize \( Q_1(\beta \mid \hat{\beta}^{(t)}) \), i.e., \( \hat{\beta}^{(t+1)} = (X^\top \hat{\Omega} X)^{-1} X^\top \hat{\Omega} y \), where \( \hat{\Omega} \) is a diagonal matrix of the weights \( \hat{\omega}_1, \ldots, \hat{\omega}_n \).

Stop when \( \| \hat{\beta}^{(t+1)} - \hat{\beta}^{(t)} \|_1 < \delta \), for a specified tolerance \( \delta > 0 \), and return the corresponding estimates of \( \beta \) and \( \psi \), as well as the weights \( \hat{\omega}_1, \ldots, \hat{\omega}_n \).

In our implementation, we initialize \( \hat{\beta}^{(1)} \) at the ordinary least squares estimator, and we take the input \( (\mathcal{W}, \psi_0, w_1, \ldots, w_n) \) for the PR portion of the algorithm as discussed in Section 3. The only adjustment required to the general PR algorithm in Section 2.1 is to add a step that calculates the weights \( \hat{\omega} \) along the sequence.

At PR–EM convergence, the corresponding weights \( \hat{\omega}_1, \ldots, \hat{\omega}_n \) become a useful by-product. In particular, these weights can be used to identify outliers. That is, when \( \hat{\omega}_i \) is close to zero, case \( i \) has little influence on the estimates of \( \beta \). So, those cases which are influential outliers will have small \( \hat{\omega} \) values, which will make them relatively easy to identify based on the PR–EM estimation. See Section 5.2 below.
To motivate our choice for the density in (5) and to justify our calling this a hybrid PR–EM algorithm, we will argue that the usual EM monotonicity property holds, i.e., if $Q_1(\beta \mid \hat{\beta}) \geq Q_1(\beta \mid \hat{\beta})$, then $\ell_{pr}(\beta) \geq \ell_{pr}(\hat{\beta})$, at least approximately. We start by rewriting $Q_2(\beta \mid \hat{\beta})$ as follows:

$$Q_2(\beta \mid \hat{\beta}) = - \sum_{i=1}^{n} \int \log \left\{ \frac{N(y_i - x_i^T \beta \mid 0, u^2)}{f_{i-1,\hat{\beta}}(y_i - x_i^T \beta)} \right\} \psi_{i,\beta}(u) \, du + nD_n(\beta, \hat{\beta}),$$

where

$$D_n(\beta, \hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_{i-1,\hat{\beta}}(y_i - x_i^T \beta)}{f_{i-1,\beta}(y_i - x_i^T \beta)}.$$

The $D_n$ term appears above because we have replaced $f_{i-1,\beta}(y_i - x_i^T \beta)$ in the denominator inside the integral with $f_{i-1,\hat{\beta}}(y_i - x_i^T \beta)$; the latter quantity is just the PR estimate $f_{i-1,\hat{\beta}}$ evaluated at $y_i - x_i^T \beta$. It is easy to see that $D_n(\beta, \hat{\beta}) = 0$, so

$$Q_2(\beta \mid \hat{\beta}) - Q_2(\hat{\beta} \mid \hat{\beta}) = \sum_{i=1}^{n} \int \log \left\{ \frac{\psi_{i,\beta}(u)}{g_{i,\beta,\hat{\beta}}(u)} \right\} \psi_{i,\beta}(u) \, du + nD_n(\beta, \hat{\beta}),$$

where

$$g_{i,\beta,\hat{\beta}}(u) = \frac{N(y_i - x_i^T \beta \mid 0, u^2) \psi_{i-1,\hat{\beta}}(u)}{f_{i-1,\hat{\beta}}(y_i - x_i^T \beta)}.$$

The integral on the inside is a Kullback–Leibler divergence and, therefore, is non-negative; it equals zero if and only if $\beta = \hat{\beta}$. Therefore, we have

$$\ell_{pr}(\beta) - \ell_{pr}(\hat{\beta}) \geq Q_1(\beta \mid \hat{\beta}) - Q_1(\hat{\beta} \mid \hat{\beta}) + nD_n(\beta, \hat{\beta}).$$

If we had that $D_n(\beta, \hat{\beta}) \geq 0$ with equality if and only if $\beta = \hat{\beta}$, then we could conclude that, by choosing $\beta$ such that $Q_1(\beta \mid \hat{\beta}) > Q_1(\hat{\beta} \mid \hat{\beta})$, one achieves $\ell_{pr}(\beta) > \ell_{pr}(\hat{\beta})$. The following heuristics explain why the inequality $D_n(\beta, \hat{\beta}) \geq 0$ should hold, at least approximately. Rewrite $D_n$ as

$$D_n(\beta, \hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f^*(y_i - x_i^T \beta)}{f_{i-1,\hat{\beta}}(y_i - x_i^T \beta)} - \frac{1}{n} \sum_{i=1}^{n} \log \frac{f^*(y_i - x_i^T \beta)}{f_{i-1,\beta}(y_i - x_i^T \beta)},$$

where $f^*$ is the true density of the errors. If we assume that $\beta$ is the true value, then the second term converges, as $n \to \infty$, to the smallest Kullback–Leibler divergence from $f^*$ over all mixtures of the specified form (Martin and Tokdar 2011). It is not clear if the first term will converge or not. If it does converge, then the limit would also be a Kullback–Leibler divergence and, by definition, cannot be smaller than the limit of the second term so, for large $n$, the difference would be non-negative. This argument is based on the assumption that $\beta$ is the true value. Therefore, the conclusion that we can reach is that if values of $\beta$ at or near the true value will increase $Q_1$ for the given $\beta$, which is intuitively quite reasonable, then we can expect that those same values will also increase $\ell_{pr}$. The obstacle to making this heuristic argument rigorous is that a theory of the behavior of the PR estimates for the non-iid case is not yet available; see Section 6.
5 Numerical results

5.1 Methods

For the numerical results in this section, we compare our hybrid PR–EM method with the following methods for robust regression; our computations are carried out using the statistical software R \cite{RCoreTeam2013}.

\begin{itemize}
  \item LS. Ordinary least squares, with the R function \texttt{lm};
  \item RLS. Robust least squares, with the default settings of the R function \texttt{rlm};
  \item ML. Maximum likelihood based on a Student-t error distribution, with df = 4, using iteratively re-weighted least squares via \texttt{lm};
  \item L\textsubscript{1}. Least absolute error regression using the default settings of the R function \texttt{rq} in the \texttt{quantreg} package \cite{Koenker2013}.
\end{itemize}

The choice of df=4 in the ML method is reasonable; see Brazzale et al. \cite[Sec. 5.2]{Brazzale2007}. R code to implement our proposed method, denoted by PREM, is available at \url{www.math.uic.edu/~rgmartin}.

5.2 Real data analysis

\textit{Example 1.} Consider a simple linear regression problem, where the predictor variable is the year, ranging from 1950 to 1973, and the response variable is the number of international phone calls from Belgium each year; so \(n = 24\). These data, available in the R software MASS library under the name \texttt{phones}, provide a classic example for robust regression \cite{Rousseeuw1987}. The scatterplot in Figure 2(a) immediately reveals the presence of several vertical outliers. The fitted lines for four methods are shown overlaid the scatter plot. An immediate conclusion is that the PREM estimate is not influenced by the outliers at all, while the vertical outliers make the other methods’ estimates (except \(L_1\)) too steep to fit the data at the later dates.

For more on PREM, we give three additional displays. First, in Figure 2(b), is a plot of the weights \(\hat{\omega}_i\) in (6). The observations with weights near 0 are exactly those apparent outliers in Figure 2(a). That these observations are assigned nearly 0 weight explains why they had essentially no influence on the fitting of the regression line. Figure 2(c) shows a plot of the PREM mixing density \(\psi\). Most of the mass is close to 0, consistent with the fact that the fluctuations around the fitted line is minimal, but there is a wide, almost imperceptible bump near \(u = 100\) which is accounting for the vertical outliers. Figure 2(d) displays the PR log-marginal likelihood path versus PR–EM iterations, and the monotonicity of the path is confirmed.

\textit{Example 2.} Here we consider an artificial data set presented in Hawkins et al. \cite{Hawkins1984}, which consists of \(n = 75\) observations and three predictor variables. This multiple regression example is considered as a benchmark for outlier detection methods. The first ten observations are regression outliers, i.e., deviations from the overall linear pattern, and the next four observations are \(x\)-outliers, or leverage points. A plot of the LS residuals versus observation number is displayed in Figure 3(a), which demonstrates the characteristics of these first 14 troublesome observations. A quantile plot of the least squares...
residuals, with simulated envelope (Atkinson 1985), is shown in Figure 3(b), which reveals the non-normality in the residuals. The PREM weights (not displayed) assigned to the four leverage points are effectively zero, so these points have no influence to the PREM fit. Figure 3(c) shows the PREM residuals, and it is clear that the PREM fit is good for all points except the four leverage points assigned weight near zero. The estimated mixing density is displayed in Figure 3(d), and it concentrates its mass in a small interval around the least squares estimator $\hat{\sigma} = 2.25$. This example shows that the PR–EM approach both removes the $x$-outliers in the model-fitting step by assigning them negligible weight and accommodates the regression outliers with a flexible model for the errors.

Example 3. In the first two examples, the presence of outliers and/or non-normality was clear; in this example, whether there is a departure from the standard Gaussian linear regression assumption is less clear. Cox and Snell (1981, Example G) present an example involving data on the construction of $n = 32$ light water reactor plants, and the mean log-cost to construct a nuclear reactor is modeled as a linear function of several predictor variables; see, also, Davison and Hinkley (1997), Brazzale et al. (2007), and Koller and Stahel (2011). After an initial variable screening, a model containing six predictor variables is considered. Two of these six predictor variables, namely $\log(N)$ and PT, which denote the number of nuclear power plants constructed by each architect–engineer...
Figure 3: Results for the Hawkins–Bradu–Kass data in Example 2. (on the log scale) and an indicator for those plants with partial turnkey guarantees, respectively, are of primary interest here. In particular, these two variables are only marginally significant based on standard regression techniques, so one could ask whether the significance of these two variables is sensitive to the choice of error distribution.

Figure 4(a) shows a quantile plot of the studentized residuals from a least squares fit, and this suggests a possibly heavier-than-normal tailed error distribution. Based on this, Brazzale et al. (2007, Sec. 5.2) employ the ML method described in Section 5.1. Figure 4(b) plots 95% confidence intervals for the slope coefficients attached to $\log(N)$ and PT based on the usual distribution theory for LS, the first-order asymptotic normality of ML, and the following method for PREM. Since $\ell_{pr}(\beta)$ is a sort of log-likelihood, Martin and Tokdar (2011) suggest that it can be used to construct confidence intervals in usual way. That is, let $J(\hat{\beta})$ denote the inverse of the Hessian matrix for $-\ell_{pr}(\beta)$ at $\beta = \hat{\beta}$, the maximizer. Then a nominal 95% confidence interval for $\beta_j$ is $\hat{\beta}_j \pm 1.96\{J(\hat{\beta})_{jj}\}^{1/2}$. In this case, the estimated coefficients for PREM and LS are almost indistinguishable, so we expect the confidence intervals to have roughly the same center. That the PREM confidence intervals are a bit longer is also to be expected since we are fitting a semiparametric model. However, given that the normal error model is reasonable, based on Figure 4(a), it is promising that the PREM intervals are not too much longer than the LS intervals. That is, it seems that PREM does not substantially
over-fit when a normal model is reasonable. Moreover, when the normality assumption is questionable, one should enlarge the model space, as PREM does, rather than change the model. When the model space is enlarged, inference should be more conservative, so we argue that the conclusions based on the PREM analysis might be more reasonable than some of those in [Brazzale et al. (2007)] and [Koller and Stahel (2011)] which are more aggressive and conclude that PT is significant.

5.3 Simulations

This section provides simulation results to compare the performance of PREM with that of the competitors listed above, under a variety of error distributions. We consider six error distributions: the two extreme exponential power distributions [West (1987)], namely, the standard normal and the standard Laplace; Student-t distributions with 1 and 2 degrees of freedom, respectively; and two non-standard normal scale mixtures, one with respect to a standard exponential distribution, denoted by N–Exp, and the other with respect to a uniform distribution supported on (0, 7), denoted by N–Unif. Both of the latter two distributions are of the general form of our model, but our support \( \mathcal{U} \) for the mixing density \( \psi \) is misspecified in both cases. Also, N–Exp has slightly heavier tails than the Laplace. These examples are far from exhaustive, but they do demonstrate that the hybrid PR–EM algorithm, with its flexible semiparametric model, is both fast and accurate compared to its competitors in a range of problems.

Example 4. In this case, we consider a regression with two predictor variables. The two predictor variables are taken to be independent standard normal samples, and the true parameter is \( \beta = 1_3 \). Table 4 gives the empirical mean square error of the estimators over 100 replications in each configuration. PREM is the only semiparametric estimator, and though it is not the best performer in all cases, it is not dominated by any other method.

Example 5. Here we consider a higher dimensional version of the simulation described in Example 4. This time we take \( p = 10 \) predictor variables, including the intercept, \( \beta = 1_{10} \), and, except for the intercept term, introduce some dependence in the predictor
variables by sampling each case from a \((p-1)\)-dimensional normal distribution with mean zero and autoregression covariance structure, with correlation parameter \(\rho = 0.5\). The same empirical mean square errors, as in Table 1, are presented in Table 2. Again, the semiparametric PREM is competitive with existing parametric methods.

### Table 1: Empirical mean squared error for the indicated method and distribution in the simulation described in Example 4.

|   | LS  | RLS | ML  | \(L_1\) | PREM  |
|---|-----|-----|-----|---------|-------|
| \(N\) | 0.029 | 0.031 | 0.033 | 0.047   | 0.037 |
| Laplace | 0.069 | 0.050 | 0.048 | 0.049   | 0.048 |
| \(t_1\) | 82.2  | 0.127 | 0.117 | 0.0924  | 0.099 |
| \(t_2\) | 0.281 | 0.065 | 0.060 | 0.068   | 0.067 |
| \(N\)-Exp | 0.060 | 0.019 | 0.018 | 0.007   | 0.008 |
| \(N\)-Unif | 0.500 | 0.317 | 0.292 | 0.212   | 0.229 |

|   | LS  | RLS | ML  | \(L_1\) | PREM  |
|---|-----|-----|-----|---------|-------|
| \(N\) | 0.171 | 0.181 | 0.185 | 0.264   | 0.255 |
| Laplace | 0.388 | 0.290 | 0.280 | 0.311   | 0.336 |
| \(t_1\) | 608  | 0.837 | 0.791 | 0.665   | 0.699 |
| \(t_2\) | 1.412 | 0.375 | 0.347 | 0.373   | 0.394 |
| \(N\)-Exp | 0.359 | 0.127 | 0.123 | 0.079   | 0.066 |
| \(N\)-Unif | 2.770 | 2.034 | 1.906 | 1.603   | 1.865 |

### Table 2: Empirical mean squared error for the indicated method and distribution in the simulation described in Example 5.

6 Discussion

This paper explores an approach to robust regression based on a semiparametric model in which the error distribution is taken to be a general scale mixture of normals with unknown mixing distribution. We estimate the regression coefficients \(\beta\) by maximizing the PR-based likelihood function \(L_{pr}(\beta)\) and, for this purpose, we propose a hybrid PR–EM algorithm based on the scale mixture of normals model for the error terms. As a by-product of the hybrid algorithm, scores are produced for each observation which can be used for outlier detection and also justify the robustness of the estimator.

The PR method in general has proved to be a useful tool in a variety of problems. However, its complicated recursive structure makes it difficult to analyze. For this reason, there remains a number of interesting open questions regarding its behavior, both asymptotics and finite samples. In particular, as we discussed above, concavity of the log-likelihood \(\ell_{pr}(\beta)\) is important for various questions about the PR estimator, i.e., existence of the estimator in finite samples, and consistency in the limit. Theoretical study of PR has been so far limited to iid models, but, as explained in Section 4.2, there is a need for further work in the independent non-iid case and beyond. We hope that this
paper can help to motivate further studies of PR both in and beyond the iid setup.

Acknowledgments

The authors are grateful to the Editor, Associate Editor, and three referees for helpful comments and suggestions, and to Byungtae Seo for providing the slides from his talk at the 2012 Joint Statistical Meetings.

References

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *J. Roy. Statist. Soc. Ser. B*, 36:99–102.

Atkinson, A. C. (1985). *Plots, Transformations, and Regression*. Oxford Univ. Press, Oxford.

Brazzale, A. R., Davison, A. C., and Reid, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics*. Cambridge University Press, Cambridge.

Carvalho, C. M., Lopez, H. F., Polson, N. G., and Taddy, M. A. (2010). Particle learning for general mixtures. *Bayesian Anal.*, 5:709–740.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.*, 90(432):1313–1321.

Cox, D. R. and Snell, E. J. (1981). *Applied Statistics*. Chapman & Hall, London.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*, volume 1. Cambridge University Press, Cambridge.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum-likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, 90(430):577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230.

Ghosh, J. K. and Tokdar, S. T. (2006). Convergence and consistency of Newton’s algorithm for estimating mixing distribution. In Fan, J. and Koul, H., editors, *Frontiers in Statistics*, pages 429–443. Imp. Coll. Press, London.

Hawkins, D. M., Bradu, D., and Kass, G. V. (1984). Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, 26(3):197–208.

Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. Unpublished manuscript, [http://www.stat.yale.edu/~pollard/Papers/convex.pdf](http://www.stat.yale.edu/~pollard/Papers/convex.pdf)
Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Statist., 1:799–821.

Huber, P. J. (1981). Robust Statistics. John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. J. Statist. Softw., 40(5):1–30.

Koenker, R. (2013). quantreg: Quantile Regression. R package version 5.05.

Koller, M. and Stahel, W. A. (2011). Sharpening Wald-type inference in robust regression for small samples. Comput. Statist. Data Anal., 55(8):2504–2515.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. J. Amer. Statist. Assoc., 84(408):881–896.

Lindsay, B. G. (1995). Mixture Models: Theory, Geometry and Applications. IMS, Hayward, CA.

Liu, C. (1996). Bayesian robust multivariate linear regression with incomplete data. J. Amer. Statist. Assoc., 91(435):1219–1227.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. Ann. Statist., 12(1):351–357.

MacEachern, S. and Müller, P. (1998). Estimating mixture of Dirichlet process models. J. Comput. Graph. Statist., 7:223–238.

Martin, R. (2013). An approximate Bayesian marginal likelihood approach for estimating finite mixtures. Comm. Statist. Simulation Comput., 42:1533–1548.

Martin, R. and Ghosh, J. K. (2008). Stochastic approximation and Newton’s estimate of a mixing distribution. Statist. Sci., 23(3):365–382.

Martin, R. and Tokdar, S. T. (2009). Asymptotic properties of predictive recursion: robustness and rate of convergence. Electron. J. Stat., 3:1455–1472.

Martin, R. and Tokdar, S. T. (2011). Semiparametric inference in mixture models with predictive recursion marginal likelihood. Biometrika, 98(3):567–582.

Martin, R. and Tokdar, S. T. (2012). A nonparametric empirical Bayes framework for large-scale multiple testing. Biostatistics, 13(3):427–439.

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. Statist. Sci., 19(1):95–110.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Statist., 9(2):249–265.

Newton, M. A. (2002). On a nonparametric recursive estimator of the mixing distribution. Sankhyā Ser. A, 64(2):306–322.
Newton, M. A., Quintana, F. A., and Zhang, Y. (1998). Nonparametric Bayes methods using predictive updating. In Dey, D., Müller, P., and Sinha, D., editors, Practical nonparametric and semiparametric Bayesian statistics, volume 133 of Lecture Notes in Statist., pages 45–61. Springer, New York.

Pinheiro, J. C., Liu, C., and Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. J. Comput. Graph. Statist., 10(2):249–276.

R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rousseeuw, P. J. and Leroy, A. M. (1987). Robust Regression and Outlier Detection. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York.

Ryan, T. P. (2009). Modern Regression Methods. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., Hoboken, NJ, second edition.

Tokdar, S. T., Martin, R., and Ghosh, J. K. (2009). Consistency of a recursive estimate of mixing distributions. Ann. Statist., 37(5A):2502–2522.

Wang, Y. (2007). On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. J. R. Stat. Soc. Ser. B, 69(2):185–198.

West, M. (1987). On scale mixtures of normal distributions. Biometrika, 74(3):646–648.