TEST (2017) 26:869–886
DOI 10.1007/s11749-017-0542-6

ORIGINAL PAPER

Robust Bayesian regression with the forward search: theory and data analysis

Anthony C. Atkinson[1] · **Aldo Corbellini**[2] · **Marco Riani**[2]

Received: 10 October 2016 / Accepted: 26 April 2017 / Published online: 11 May 2017

Abstract The frequentist forward search yields a flexible and informative form of robust regression. The device of fictitious observations provides a natural way to include prior information in the search. However, this extension is not straightforward, requiring weighted regression. Bayesian versions of forward plots are used to exhibit the presence of multiple outliers in a data set from banking with 1903 observations and nine explanatory variables which shows, in this case, the clear advantages from including prior information in the forward search. Use of observation weights from frequentist robust regression is shown to provide a simple general method for robust Bayesian regression.

Keywords Consistency factor · Fictitious observation · Forward search · Graphical methods · Outliers · Weighted regression

Mathematics Subject Classification 62F15 · 62F35 · 62J05 · 65C60 · 68U20

1 Introduction

Frequentist methods for robust regression are increasingly studied and applied. The foundations of robust statistical methods are presented in the books of Hampel et al.

✉ Anthony C. Atkinson
a.c.atkinson@lse.ac.uk

Aldo Corbellini
aldo.corbellini@unipr.it

Marco Riani
mriani@unipr.it

[1] The London School of Economics, London WC2A 2AE, UK

[2] Dipartimento di Scienze Economiche e Aziendali, Università di Parma, Parma, Italy

(1986), of Maronna et al. (2006) and of Huber and Ronchetti (2009). Book length treatments of robust regression include Rousseeuw and Leroy (1987) and Atkinson and Riani (2000). However, none of these methods includes prior information; they can all be thought of as robust developments of least squares. The present paper describes a procedure for robust regression incorporating prior information, determines its properties and illustrates its use in the analysis of a dataset with 1903 observations.

The purpose of the robust analysis is to detect outlying observations; these may be isolated or form clusters, or indicate systematic departures from the fitted model. Once the outliers have been downweighted or deleted, there remains a set of "clean" data, in agreement with the fitted model. It is helpful to divide frequentist methods of robust regression into three classes.

1. Hard (0,1) Trimming. In Least Trimmed Squares (LTS: Hampel 1975; Rousseeuw 1984) the amount of trimming of the n observations when the number of parameters in the full-rank model is p, is determined by the choice of the trimming parameter h, $[n/2] + [(p+1)/2] \leq h \leq n$, which is specified in advance. The LTS estimate is intended to minimize the sum of squares of the residuals of h observations. In Least Median of Squares (LMS: Rousseeuw 1984) the estimate minimizes the median of h squared residuals.
2. Adaptive Hard Trimming. In the Forward Search (FS), the observations are again hard trimmed, but the value of h is determined by the data, being found adaptively by the search. Data analysis starts from a very robust fit to a few, carefully selected, observations found by LMS or LTS with the minimum value of h. The number of observations used in fitting then increases until all are included.
3. Soft Trimming (downweighting). M estimation and derived methods (Huber and Ronchetti 2009). The intention is that observations near the centre of the distribution retain their value, but the function ρ, which determines the form of trimming, ensures that increasingly remote observations have a weight that decreases with distance from the centre.

We use the Forward Search as the basis for our proposed method of robust Bayesian regression. Other methods, such as LTS, S or MM, included in the comparisons of Riani et al. (2014c) of frequentist methods of robust regression, could also be extended to provide robust procedures incorporating prior information. In Sect. 6 we briefly indicate one method of doing this.

As we describe in more detail in the next section, the FS uses least squares to fit the model to subsets of m observations, chosen to have the m smallest squared residuals, the subset size increasing during the search. The results of the FS are typically presented through a forward plot of quantities of interest as a function of m. As a result, it is possible to connect individual observations with changes in residuals and parameter estimates, thus identifying outliers and systematic failures of the fitted model. (See Atkinson et al. (2010) for a general survey of the FS, with discussion). In addition, since the method is based on the repeated use of least squares, it is relatively straightforward to introduce prior information into the search.

Whichever of the three forms of robust regression given above is used, the aim in outlier detection is to obtain a "clean" set of data providing estimates of the parameters uncorrupted by any outliers. Inclusion of outlying observations in the data subset used

for parameter estimation can yield biased estimates of the parameters, making the outliers seem less remote, a phenomenon called "masking". The FS avoids masking by the use, for as large a value of m as possible, of observations believed not to be outlying. The complementary "backward" procedure starts with diagnostic measures calculated from all the data and then deletes the most outlying. The procedure continues until no further outliers are identified. Such procedures, described in the books of Cook and Weisberg (1982) and Atkinson (1985), are prone to the effect of masking. Illustrations of this effect for several different models are in Atkinson and Riani (2000) and demonstrate the failure of the method to identify outliers. The Bayesian outlier detection methods of West (1984) and Chaloner and Brant (1988) start from parameter estimates from the full sample and so can also be expected to suffer from masking.

Although it is straightforward to introduce prior information into the FS, an interesting technical problem arises in estimation of the error variance σ^2. Since the sample estimate in the frequentist search comes from a set of order statistics of the residuals, the estimate of σ^2 has to be rescaled. In the Bayesian search, we need to combine a prior estimate with one obtained from such a set of order statistics from the subsample of observations. This estimate has likewise to be rescaled before being combined with the prior estimate of σ^2; parameter estimation then uses weighted least squares. A similar calculation could be used to provide a version of least trimmed squares (Rousseeuw 1984) that incorporates prior information. Our focus throughout is on linear regression, but our technique of representing prior information by fictitious observations can readily be extended to more complicated models such as those based on ordinal regression described in Croux et al. (2013) or for sparse regression (Hoffmann et al. 2015).

The paper is structured as follows. Notation and parameter estimation for Bayesian regression are introduced in Sect. 2. Section 2.3 describes the introduction into the FS of prior information in the form of fictitious observations, leading to a form of weighted least squares which is central to our algorithm. We describe the Bayesian FS in Sect. 3 and, in Sect. 4, use forward plots to elucidate the change in properties of the search with variation of the amount of prior information. The example, in Sect. 5, shows the effect of the outliers on parameter estimation and a strong contrast with the frequentist analysis which indicated over twelve times as many outliers. In Sect. 6 a comparison of the forward search with a weighted likelihood procedure (Agostinelli 2001) leads to a general method for the extension of robust frequentist regression to include prior information. A simulation study in Sect. 7 compares the power of frequentist and Bayesian procedures, both when the prior specification is correct and when it is not. The paper concludes with a more general discussion in Sect. 8.

2 Parameter estimation

2.1 No prior information

We first, to establish notation, consider parameter estimation in the absence of prior information, that is least squares.

In the regression model $y = X\beta + \varepsilon$, y is the $n \times 1$ vector of responses, X is an $n \times p$ full-rank matrix of known constants, with ith row x_i^{T}, and β is a vector of p

unknown parameters. The normal theory assumptions are that the errors ε_i are i.i.d. $N(0, \sigma^2)$.

The least squares estimator of β is $\hat{\beta}$. Then the vector of n least squares residuals is $e = y - \hat{y} = y - X\hat{\beta} = (I - H)y$, where $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is the 'hat' matrix, with diagonal elements h_i and off-diagonal elements h_{ij}. The residual mean square estimator of σ^2 is $s^2 = e^{\mathrm{T}}e/(n-p) = \sum_{i=1}^{n} e_i^2/(n-p)$.

In order to detect outliers and departures from the fitted regression model, FS uses least squares to fit the model to subsets of m observations. The subset is increased from size m to $m+1$ by forming the new subset from the observations with the $m+1$ smallest squared residuals. For each m $(m_0 \le m \le n-1)$, we use deletion residuals to test for the presence of outliers. These tests require an estimate of σ^2. If we estimated σ^2 from all n observations, the statistics would have a t distribution on $n-p$ degrees of freedom. However, in the search we select the central m out of n observations to provide the estimate $s^2(m)$, so that the variability is underestimated. To allow for estimation from this truncated distribution, let the variance of the symmetrically truncated standard normal distribution containing the central m/n portion of the full distribution be

$$c(m,n) = 1 - \frac{2n}{m}\Phi^{-1}\left(\frac{n+m}{2n}\right)\phi\left\{\Phi^{-1}\left(\frac{n+m}{2n}\right)\right\}, \tag{1}$$

where $\phi(.)$ and $\Phi(.)$ are, respectively, the standard normal density and c.d.f. See Riani et al. (2009) for a derivation from the general method of Tallis (1963). We take $s^2(m)/c(m,n)$ as our approximately unbiased estimate of variance. In the robustness literature, the important quantity $c(m,n)$ is called a consistency factor (Riani et al. 2014b; Johansen and Nielsen 2016).

2.2 The normal inverse-gamma prior distribution

We represent prior information using the conjugate prior for the normal theory regression model leading to a normal prior distribution for β and an inverse-gamma distribution for σ^2.

If the density of the gamma distribution $G(a,b)$ is written

$$f_G(x,a,b) = \frac{b^a}{\Gamma(a)}x^{a-1}\exp(-bx),$$

$G(a,b)$ has mean a/b and variance a/b^2.

If $X \sim G(a,b)$, then $Y = 1/X$ has an inverse-gamma distribution $IG(a,b)$ with density

$$f_{IG}(x,a,b) = \frac{b^a}{\Gamma(a)}(1/x)^{a+1}\exp(-b/x) \quad (x > 0),$$

shape parameter a and scale parameter b. The mean (for $a > 1$) is $b/(a-1)$, and the variance (for $a > 2$) is $b^2/(a-1)^2(a-2)$.

Let the values of the parameters specifying the prior distribution be a_0, b_0, β_0 and R. Then the normal inverse-gamma conjugate family of prior distributions for β and σ^2 has the form

$$f(\beta, \sigma^2) \propto (1/\sigma^2)^{a_0+1+\frac{p}{2}} \exp\left\{-\frac{(\beta-\beta_0)^{\mathrm{T}} R(\beta-\beta_0)}{2\sigma^2} - \frac{b_0}{\sigma^2}\right\}.$$

The marginal distribution of σ^2 is $IG(a_0, b_0)$. Let $\tau = 1/\sigma^2$. Then $f(\tau) \propto \tau^{a_0-1} \exp(-b_0\tau)$, that is $G(a_0, b_0)$. The prior distribution of β conditional on τ is $N\{\beta_0, (1/\tau)R^{-1}\}$.

2.3 Prior distribution from fictitious observations

The device of fictitious prior observations provides a convenient representation of this conjugate prior information. We follow, for example, Chaloner and Brant (1988), who are interested in outlier detection, and describe the parameter values of these prior distributions in terms of n_0 fictitious observations.

We start with σ^2. Let the estimate of σ^2 from the n_0 fictitious observations be s_0^2 on $n_0 - p$ degrees of freedom. Then in $f(\tau)$,

$$a_0 = \nu_0/2 = (n_0 - p)/2 \quad \text{and} \quad b_0 = \nu_0 s_0^2/2 = S_0/2,$$

where S_0 is the residual sum of squares of the fictitious observations.

Prior information for the linear model is given as the scaled information matrix $R = X_0^{\mathrm{T}} X_0$ and the prior mean $\hat{\beta}_0 = R^{-1} X_0^{\mathrm{T}} y_0$. Then $S_0 = y_0^{\mathrm{T}} y_0 - \hat{\beta}_0^{\mathrm{T}} R \hat{\beta}_0$. Thus, given n_0 prior observations the parameters for the normal inverse-gamma prior may readily be calculated.

2.4 Posterior distributions

The posterior distribution of β conditional on τ is $N\{\hat{\beta}_1, (1/\tau)(R + X^{\mathrm{T}}X)^{-1}\}$ where

$$\begin{aligned}\hat{\beta}_1 &= (R + X^{\mathrm{T}}X)^{-1}(R\beta_0 + X^{\mathrm{T}}y) \\ &= (R + X^{\mathrm{T}}X)^{-1}(R\beta_0 + X^{\mathrm{T}}X\hat{\beta}) \\ &= (I - A)\beta_0 + A\hat{\beta}, \end{aligned} \tag{2}$$

and $A = (R + X^{\mathrm{T}}X)^{-1} X^{\mathrm{T}}X$. The last expression shows that the posterior estimate $\hat{\beta}_1$ is a matrix weighted average of the prior mean β_0 and the classical OLS estimate $\hat{\beta}$, with weights $I - A$ and A. If prior information is strong, the elements of R will be large, and A will be small, so that the posterior mean gives most weight to the prior mean. In the classical approach these weights are fixed, while with the forward search, as the subset size grows, the weight assigned to A increases with m; we can dynamically see how the estimate changes as the effect of the prior decreases.

The posterior distribution of τ is $G(a_1, b_1)$ where

$$a_1 = a + n/2 = (n_0 + n - p)/2 \quad \text{and} \tag{3}$$

$$b_1 = \left\{(n_0 - p)/\tau_0 + (y - X\beta_1)^{\mathrm{T}} y + (\beta_0 - \beta_1)^{\mathrm{T}} R\beta_0\right\}/2. \tag{4}$$

The posterior distribution of σ^2 is $IG(a_1, b_1)$. The posterior mean estimates of τ and σ^2 are, respectively,

$$\tau_1 = a_1/b_1, \quad \text{and} \quad \tilde{\sigma}_1^2 = b_1/(a_1 - 1). \tag{5}$$

In our calculations, we take $\hat{\sigma}_1^2 = 1/\tau_1$ as the estimate of σ^2. Unless a_1 is very small, the difference between $\hat{\sigma}_1^2$ and $\tilde{\sigma}_1^2$ is negligible.

The posterior marginal distribution of β is multivariate t with parameters

$$\hat{\beta}_1, \ (1/\tau_1)\{R + X^{\mathrm{T}}X\}^{-1}, \ n_0 + n - p.$$

3 The Bayesian search

3.1 Parameter estimation

The posterior distributions of Sect. 2.4 arise from the combination of n_0 prior observations, perhaps fictitious, and the n actual observations. In the FS we combine the n_0 prior observations with a carefully selected m out of the n observations. The search proceeds from $m = 0$, when the fictitious observations provide the parameter values for all n residuals from the data. It then continues with the fictitious observations always included amongst those used for parameter estimation; their residuals are ignored in the selection of successive subsets.

As mentioned in Sect. 1, there is one complication in this procedure. The n_0 fictitious observations are treated as a sample with population variance σ^2. However, the m observations from the actual data are, as in Sect. 2.1, from a truncated distribution of m out of n observations and so asymptotically have a variance $c(m, n)\sigma^2$. An adjustment must be made before the two samples are combined. This becomes a problem in weighted least squares (for example, Rao 1973, p. 230). Let y^+ be the $(n_0 + m) \times 1$ vector of responses from the fictitious observations and the subset, with X^+ the corresponding matrix of explanatory variables. The covariance matrix of the independent observations is $\sigma^2 G$, with G a diagonal matrix; the first n_0 elements of the diagonal of G equal one and the last m elements have the value $c(m, n)$. The information matrix for the $n_0 + m$ observations is

$$(X^{+\mathrm{T}} W X^+)/\sigma^2 = \{X_0^{\mathrm{T}} X_0 + X(m)^{\mathrm{T}} X(m)/c(m, n)\}/\sigma^2, \tag{6}$$

where $W = G^{-1}$. In the least squares calculations, we need only to multiply the elements of the sample values $y(m)$ and $X(m)$ by $c(m, n)^{-1/2}$. However, care is needed to obtain the correct expressions for leverages and variances of parameter estimates.

Since, during the forward search, n in (3) is replaced by the subset size m, X and y in (4) become $y(m)/\sqrt{c(m, n)}$ and $X(m)/\sqrt{c(m, n)}$, giving rise to posterior values $a_1(m)$, $b_1(m)$, $\tau_1(m)$ and $\hat{\sigma}_1^2(m)$.

The estimate of β from n_0 prior observations and m sample observations can, from (6), be written

$$\hat{\beta}_1(m) = (X^{+\mathrm{T}} W X^+)^{-1} X^{+\mathrm{T}} W y^+. \tag{7}$$

In Sect. 2.3 $\hat{\beta}_0 = R^{-1} X_0^{\mathrm{T}} y_0$, so that $X_0^{\mathrm{T}} y_0 = X_0^{\mathrm{T}} X_0 \hat{\beta}_0$. Then the estimate in (7) can be written in full as

$$\begin{aligned} \hat{\beta}_1(m) &= \{X_0^{\mathrm{T}} X_0 + X(m)^{\mathrm{T}} X(m)/c(m,n)\}^{-1} \{X_0^{\mathrm{T}} y_0 + X(m)^{\mathrm{T}} y(m)/c(m,n)\} \\ &= \{X_0^{\mathrm{T}} X_0 + X(m)^{\mathrm{T}} X(m)/c(m,n)\}^{-1} \{X_0^{\mathrm{T}} X_0 \hat{\beta}_0 + X(m)^{\mathrm{T}} y(m)/c(m,n)\}. \end{aligned} \tag{8}$$

3.2 Forward highest posterior density intervals

Inference about the parameters of the regression model comes from regions of highest posterior density. These are calculated from the prior information and the subset at size m. Let

$$V(m) = (X^{+\mathrm{T}} X^+)^{-1} = \{X_0^{\mathrm{T}} X_0 + X(m)^{\mathrm{T}} X(m)\}^{-1}, \tag{9}$$

with (j, j)th element $V_{jj}(m)$. Likewise, the j-th element of $\hat{\beta}_1(m)$, $j = 1, 2, \ldots, p$ is denoted $\hat{\beta}_{1j}(m)$. Then

$$\operatorname{var} \hat{\beta}_{1j}(m) = \hat{\sigma}^2(m) V_{jj}(m).$$

The $(1-\alpha)\%$ highest posterior density (HPD) interval for β_{1j} is

$$\hat{\beta}_{1j}(m) \pm t_{\nu,1-\alpha/2} \sqrt{\hat{\sigma}^2(m) V_{jj}},$$

with $t_{\nu,\gamma}$ the $\gamma\%$ point of the t distribution on ν degrees of freedom. Here $\nu = n_0 + m - p$.

The highest posterior density intervals for τ and σ^2 are, respectively, given by

$$[g_{a_1(m),b_1(m),\alpha/2}, g_{a_1(m),b_1(m),1-\alpha/2}] \quad \text{and} \quad [ig_{a_1(m),b_1(m),\alpha/2}, ig_{a_1(m),b_1(m),1-\alpha/2}],$$

where $g_{a,b,\gamma}$ and $ig_{a,b,\gamma}$ are the $\gamma\%$ points of the $G(a,b)$ and $IG(a,b)$ distributions.

3.3 Outlier detection

We detect outliers using a form of deletion residual that includes the prior information. Let $S^*(m)$ be the subset of size m found by FS, for which the matrix of regressors is $X(m)$. Weighted least squares on this subset of observations (8) yields parameter estimates $\hat{\beta}_1(m)$ and $\hat{\sigma}^2(m)$, an estimate of σ^2 on $n_0 + m - p$ degrees of freedom. The residuals for all n observations, including those not in $S^*(m)$, are

$$e_i(m) = y_i - x_i^{\mathrm{T}} \hat{\beta}_1(m) \qquad (i = 1, \ldots, n). \tag{10}$$

The search moves forward with the augmented subset $S^*(m+1)$ consisting of the observations with the $m+1$ smallest absolute values of $e_i(m)$. To start we take $m_0 = 0$, since the prior information specifies the values of β and σ^2.

To test for outliers, the deletion residuals are calculated for the $n-m$ observations not in $S^*(m)$. These residuals are

$$r_i(m) = \frac{y_i - x_i^{\mathrm{T}}\hat{\beta}_1(m)}{\sqrt{\hat{\sigma}^2(m)\{1+h_i(m)\}}} = \frac{e_i(m)}{\sqrt{\hat{\sigma}^2(m)\{1+h_i(m)\}}}, \tag{11}$$

where, from (8), the leverage $h_i(m) = x_i^{\mathrm{T}}\{X_0^{\mathrm{T}}X_0 + X(m)^{\mathrm{T}}X(m)/c(m,n)\}^{-1}x_i$. Let the observation nearest to those forming $S^*(m)$ be $i_{\min}$ where

$$i_{\min} = \arg\min_{i \notin S^*(m)} |r_i(m)|.$$

To test whether observation $i_{\min}$ is an outlier, we use the absolute value of the minimum deletion residual

$$r_{\mathrm{imin}}(m) = \frac{e_{\mathrm{imin}}(m)}{\sqrt{\hat{\sigma}^2(m)\{1+h_{\mathrm{imin}}(m)\}}}, \tag{12}$$

as a test statistic. If the absolute value of (12) is too large, the observation $i_{\min}$ is considered to be an outlier, as well as all other observations not in $S^*(m)$.

3.4 Envelopes and multiple testing

A Bayesian FS through the data provides a set of n absolute minimum deletion residuals. We require the null pointwise distribution of this set of values and find, for each value of m, a numerical estimate of, for example, the 99% quantile of the distribution of $|r_{\mathrm{imin}}(m)|$.

When used as the boundary of critical regions for outlier testing, these envelopes have a pointwise size of 1%. Performing n tests of outlyingness of this size leads to a procedure for the whole sample which has a size much greater than the pointwise size. In order to obtain a procedure with a 1% samplewise size, we require a rule which allows for the simple behaviour in which a few outliers enter at the end of the search and the more complicated behaviour when there are many outliers which may be apparent away from the end of the search. However, at the end of the search such outliers may be masked and not evident. Our chosen rule achieves this by using exceedances of several envelopes to give a "signal" that outliers may be present.

In cases of appreciable contamination, the signal may occur too early, indicating an excessive number of outliers. This happens because of the way in which the envelopes increase towards the end of the search. Accordingly, we check the sample size indicated by the signal for outliers and then increase it, checking the 99% envelope for outliers as the value of n increases, a process known as resuperimposition. The notation $r_{\min}(m,n)$ indicates the dependence of this process on a series of values of n.

In the next section, where interest is in envelopes over the whole search, we find selected percentage points of the null distribution of $|r_{\mathrm{imin}}(m)|$ by simulation. However,

in the data analyses of Sect. 5 the focus is on the detection of outliers in the second half of the search. Here we use a procedure derived from the distribution of order statistics to calculate the envelopes for the many values of $r_{\min}(m, n)$ required in the resuperimposition of envelopes. Further details of the algorithm and its application to the frequentist analysis of multivariate data are in Riani et al. (2009).

4 Prior information and simulation envelopes

We now illustrate the effect of prior information on the envelopes. Figure 1 shows the results of 10,000 simulations of normally distributed observations from a regression model with four variables and a constant ($p = 5$), the values of the explanatory variables having independent standard normal distributions. These envelopes are invariant to the numerical values of β and σ^2. The left-hand panel shows 1, 50 and 99% simulation envelopes for weak prior information when $n_0 = 30$ (and $n = 500$), along with the envelopes in the absence of any prior information. As m increases the two sets of envelopes become virtually indistinguishable, illustrating the irrelevance of this amount of prior information for such large samples. On the other hand, the right-hand panel keeps $n = 500$, but now n_0 has the same value. There is again good agreement between the two sets of envelopes towards the end of the search, especially for the upper envelope.

In our example in Sect. 5, we not only look at outlier detection, but also at parameter estimation. The left-hand panel of Fig. 2 shows empirical quantiles for the distribution of $\hat{\beta}_3(m)$ from 10,000 simulations when $\beta_3 = 0$. Because of the symmetry of our simulations, this is indistinguishable from the plots for the other parameters of the linear model. The right-hand panel shows the forward plot of $\hat{\sigma}^2(m)$, simulated with $\sigma^2 = 1$. In this simulation the prior information, with $n_0 = 30$, is small compared with the sample information. In the forward plot for $\hat{\beta}_3$ the bands are initially wide, but rapidly narrow, being symmetrical about the simulation value of zero. There are two

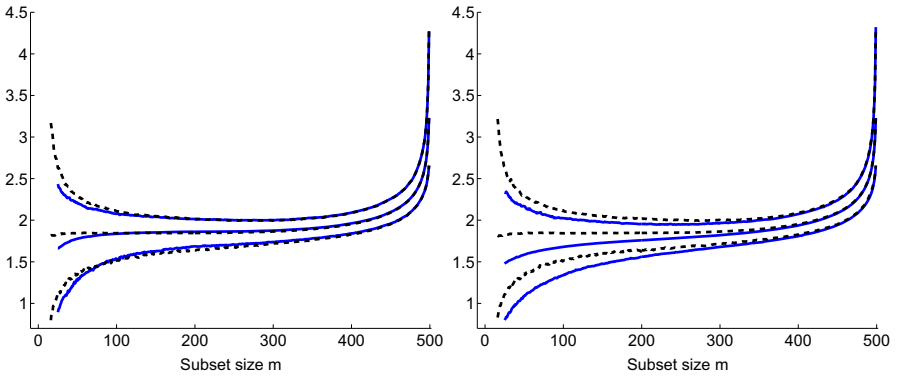

Fig. 1 The effect of correct prior information on forward plots of envelopes of absolute Bayesian minimum deletion residuals. *Left-hand panel*, weak prior information ($n_0 = 30$; $n = 500$). *Right-hand panel*, strong prior information ($n_0 = 500$; $n = 500$), 10,000 simulations; 1, 50 and 99% empirical quantiles. *Dashed lines*, without prior information; *heavy lines*, with prior information

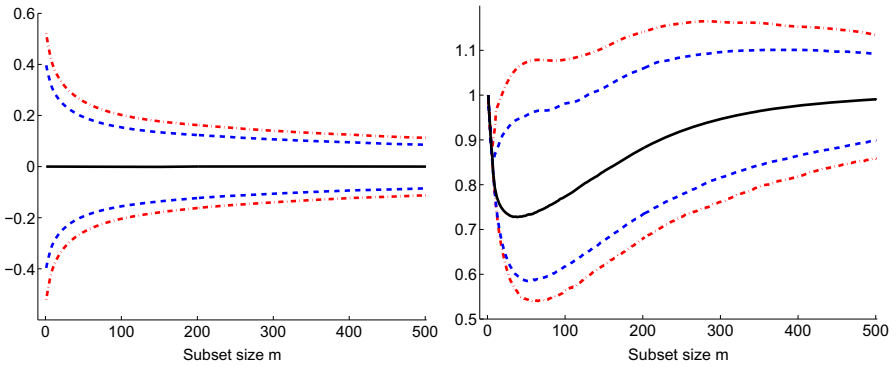

Fig. 2 Distribution of parameter estimates when $\beta_3 = 0$ and $\sigma^2 = 1$. *Left-hand panel* $\hat{\beta}_3(m)$, *right-hand panel* $\hat{\sigma}^2(m)$; weak prior information ($n_0 = 30$; $n = 500$). 1, 5, 50, 95 and 99% empirical quantiles

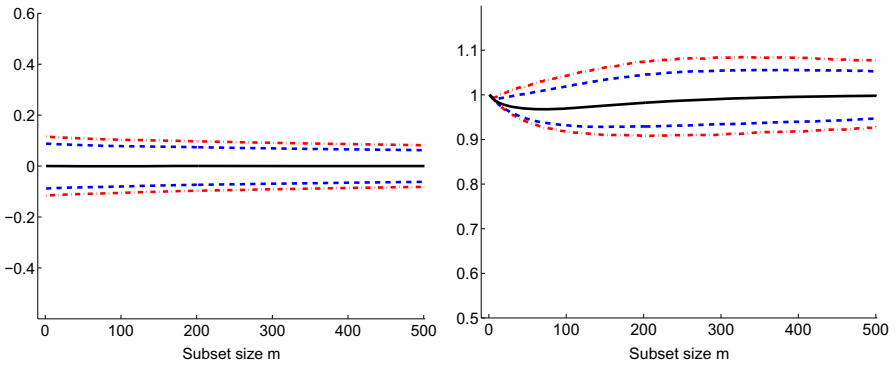

Fig. 3 Distribution of parameter estimates. *Left-hand panel* $\hat{\beta}_3(m)$, *right-hand panel* $\hat{\sigma}^2(m)$; strong prior information ($n_0 = 500$; $n = 500$). 1, 5, 50, 95 and 99% empirical quantiles. *Vertical scale* the same as that of Fig. 2

effects causing the initial rapid decrease in the width of the interval during the FS. The first is under-estimation of σ^2 which, as the right-hand panel shows, has a minimum value around 0.73. This under-estimation occurs because $c(m, n)$ is an asymptotic correction factor. Further correction is needed in finite samples. Pison et al. (2002) use simulation to make such corrections in robust regression, but not for FS. The second effect is again connected with the value of $c(m, n)$, which is small for small m/n (for example 0.00525 for 10%). Then, from (6), the earliest observations to enter the search will have a strong effect on reducing var $\hat{\beta}(m)$.

The panels of Fig. 3 are for similar simulations, but now with n_0 and n both 500. The main differences from Fig. 2 are that the widths of the bands now decrease only slightly with m and that the estimate of σ^2 is relatively close to one throughout the search; the minimum value in this simulation is 0.97.

The widths of the intervals for $\hat{\beta}_3(m)$ depend on the information matrices. If, as here, the prior data and the observations come from the same population, the ratio of the widths of the prior band to that at the end of the search is $\sqrt{\{(n_0+n-p)/(n_0-p)\}}$,

Table 1 Bank profit data: prior estimates of parameters

Parameter	β_0	β_1	β_2	β_3	β_4	β_5
Mean	−0.5	9.1	0.001	0.0002	0.002	0.12
Parameter	β_6	β_7	β_8	β_9	s_0^2	
Mean	0.0004	−0.0004	1.3	0.00004	10,000	

here $\sqrt{(525/25)}$, or approximately 4.58, for the results plotted in Fig. 2. In Fig. 3 the ratio is virtually $\sqrt{2}$. This difference is clearly reflected in the figures.

5 Example: bank profit data

As an example of the application of the Bayesian FS, we now analyse data on the profitability to an Italian bank of customers with a variety of profiles, as measured by nine explanatory variables.

The data are the annual profit from 1903 customers, all of whom were selected by the bank as the target for a specific campaign. The data are available in the FSDA toolbox under the title BankProfit. The nine explanatory variables are either amounts at a particular time point, or totals over the year. Together with the response they are:

- y_i: annual profit or loss per customer;
- x_{1i}: number of products bought by the customers;
- x_{2i}: current account balance plus holding of bonds issued by the bank;
- x_{3i}: holding of investments for which the bank acted as an agent;
- x_{4i}: amount in deposit and savings accounts with the bank;
- x_{5i}: number of activities in all accounts;
- x_{6i}: total value of all transactions;
- x_{7i}: total value of debit card spending (recorded with a negative sign);
- x_{8i}: number of credit and debit cards;
- x_{9i}: total value of credit card spending.

The prior values of the eleven parameters, directly supplied by the bank, are given in Table 1. The values of n_0 and a_0 are 1500 and 745, appreciable compared to the 1903 observations; $b_0 = 7{,}450{,}000$. The matrix R is 10×10 and is therefore only given in the toolbox. Apart from the intercept and β_7, all parameter values are positive. However the values of x_7 are recorded as negative values, so that profit is expected to increase with large negative values of the variable.

The prior estimates of the parameters come from a non-robust analysis of earlier data. The purpose of the present analysis is to see what are the most striking changes in the importance of the variables for predicting profitability when a robust analysis is used which removes masked outliers and their associated effects on parameter estimates.

Figure 4 shows the forward plot of absolute Bayesian deletion residuals from $m = 1700$. There is a signal at $m = 1763$. However, the use of resuperimposition of

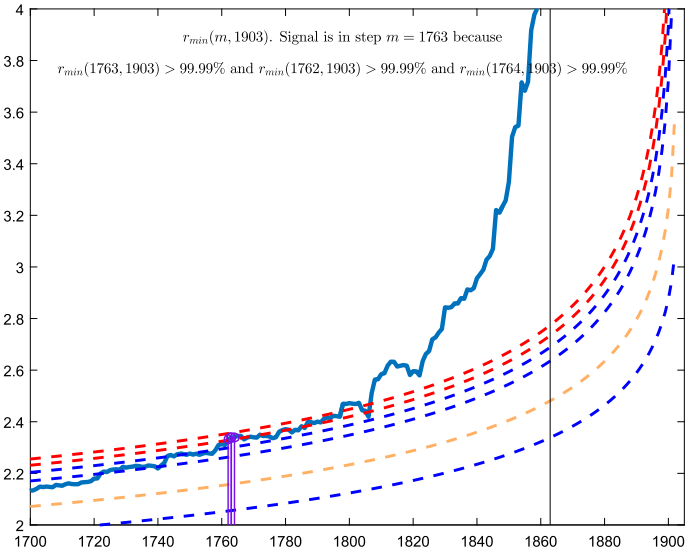

Fig. 4 Bank profit data; forward plot of absolute Bayesian minimum deletion residuals. There is a signal indicating outliers from $m = 1763$. Prior information as in Table 1

envelopes leads to the identification of 48 outliers; the signal occurs at a smaller value of m than would be expected from the number of outliers finally identified.

Scatter plots of the data, showing the outliers, are in Fig. 5. The figure shows there are eleven exceptionally profitable customers and three exceptionally unprofitable ones. The data for these individuals should clearly be checked to determine whether they appear so exceptional due to data errors. Otherwise, the observations mostly fall in clusters or around lines, although further outliers are generated by anomalously high values of some of the explanatory variables. The main exception is x_4 where the outliers show as a vertical line in the plot, distinct from the strip containing the majority of the observations.

Figure 6 shows the forward plots of the HPD regions, together with 95 and 99% envelopes. The horizontal lines indicate the prior values of the parameters and the vertical line indicates the point at which outliers start to be included in the subset used for parameter estimation.

These results show very clearly the effect of the outliers. In the left-hand part of the panels and, indeed, in the earlier part of the search not included in the figure, the parameter estimates are stable, in most cases lying close to their prior values. However, inclusion of the outliers causes changes in the estimates. Some, such as $\hat{\beta}_1(m)$, $\hat{\beta}_3(m)$ and $\hat{\beta}_7(m)$, move steadily in one direction. Others, such as $\hat{\beta}_6(m)$ and $\hat{\beta}_9(m)$, oscillate, especially towards the very end of the search. The most dramatic change is in $\hat{\beta}_4(m)$ which goes from positive to negative as the vertical strip of outliers is included. From a banking point of view, the most interesting results are those for the two parameters with negative prior values. It might be expected that the intercept would be zero or slightly negative. But $\hat{\beta}_7(m)$ remains positive throughout the search, thus changing understanding of the importance of x_7, debit card spending. More generally important

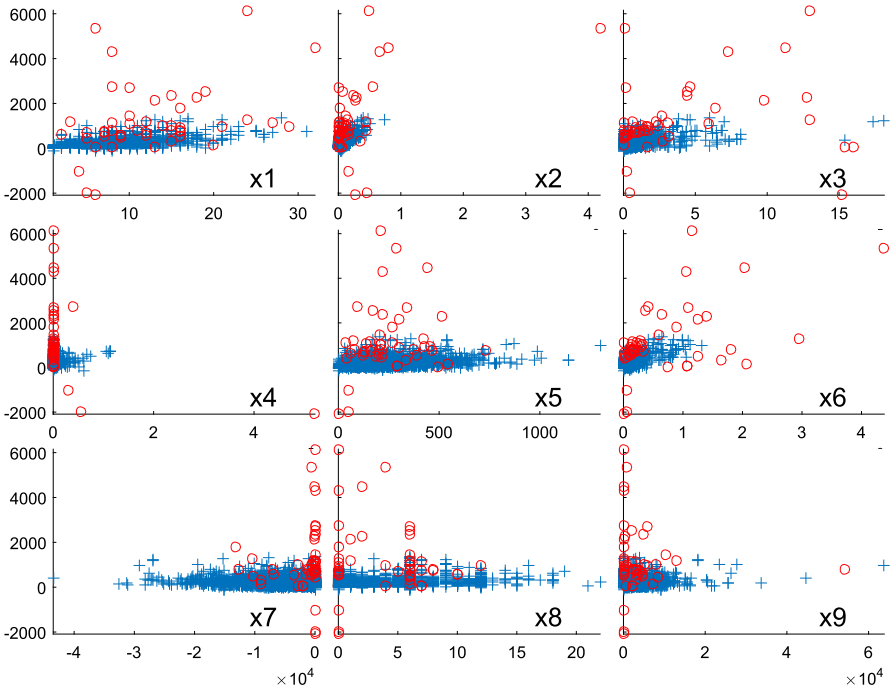

Fig. 5 Bank profit data; scatter plots of y against the nine x variables, indicating the outliers found by the Bayesian FS. Outliers ∘, other observations +

is the appreciable increase in the estimate of σ^2. In the figure this has been truncated, so that the stability of the estimate in the earlier part of the search is visible. However, when all observations are used in fitting, the estimate has a value of 3.14e+04, as opposed to a value close to 1.0e+04 for much of the outlier free search. Such a large value renders inferences imprecise, with some loss of information. This shows particularly clearly in the plots of those estimates less affected by outliers, such as $\hat{\beta}_0(m)$, $\hat{\beta}_5(m)$ and $\hat{\beta}_8(m)$.

The 95 and 99% HPD regions in Fig. 6 also provide information about the importance of the predictors in the model. In the absence of outliers, only the regions for $\hat{\beta}_0(m)$, $\hat{\beta}_8(m)$ and $\hat{\beta}_9(m)$ include zero, so that these terms might be dropped from the model, although dropping one term might cause changes in the HPD regions for the remaining variables. The effect of the outliers is to increase the seeming importance of some other variables, such as x_1 and x_3. Only $\hat{\beta}_4(m)$ shows a change of sign.

We do not make a detailed comparison with the frequentist forward search which declares 586 observations as outliers. This apparent abundance of outliers is caused by anomalously high values of some of the explanatory variables. Such high leverage points can occasionally cause misleading fluctuations in the forward search trajectory leading to early stopping. However, such behaviour can be detected by visual inspection of such plots as the frequentist version of Fig. 4. The Bayesian analysis provides a stability in the procedure which avoids an unnecessary rejection of almost one third of the data.

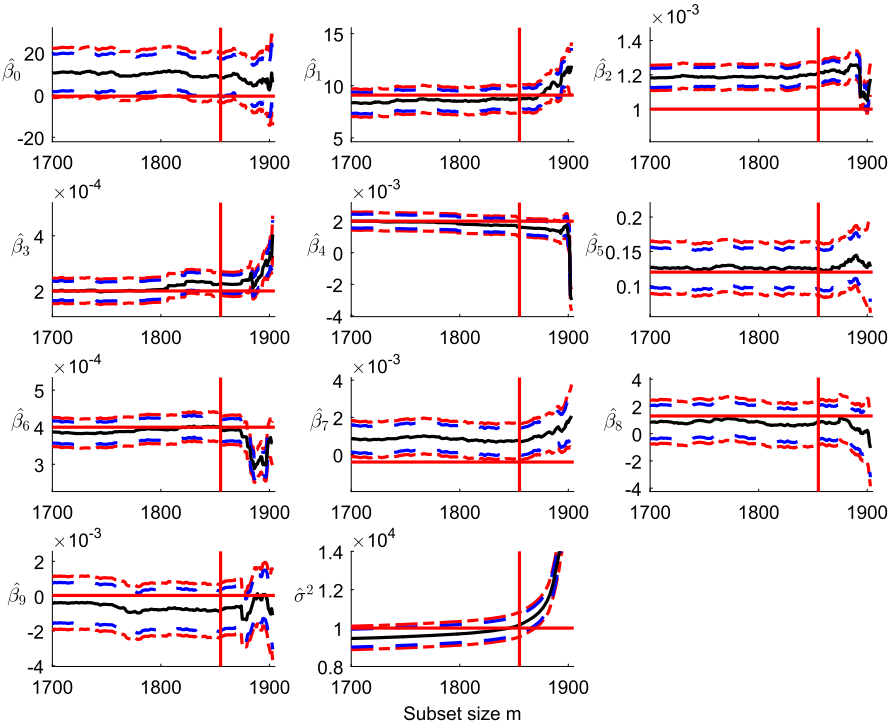

Fig. 6 Bank profit data; forward plots of 95 and 99% HPD regions for the parameters of the linear model and, *bottom right-hand panel*, the estimate of σ^2. The last part of the search from $m = 1700$. The *horizontal lines* correspond to prior values, the *vertical line* to the point at which outliers start to enter the subset

6 A comparison with weighted likelihood

6.1 Background

The fundamental output of a robust analysis is the weight attached to each observation. In the forward search, the adaptively calculated weights have the values 0 and 1; in the analysis of the bank profit data the weights from the forward search contain 48 zeroes.

Many other robust methods, such as MM- and S-estimation (Maronna et al. 2006), downweight observations in a more smooth way, resulting in weights that have values in [0,1]. As an example, we use the trimmed likelihood weights from the R package `wle` (Agostinelli 2001). The calculation of these robust weights, which forms a first stage of their Bayesian analysis, is described in Agostinelli and Greco (2013, §2). Incorporation of prior information forms a second stage.

Once the output of a robust analysis is viewed as a set of weights, it is straightforward to incorporate prior information into the analysis using the results on parameter estimation from Sect. 3.1. In particular, the posterior estimate of the vector of parameters in the linear model follows immediately from (8) as

$$\begin{aligned}\hat{\beta}_1 &= \{X_0^{\mathrm{T}} X_0 + X^{\mathrm{T}} W_R X\}^{-1} \{X_0^{\mathrm{T}} y_0 + X^{\mathrm{T}} W_R y\} \\ &= \{X_0^{\mathrm{T}} X_0 + X^{\mathrm{T}} W_R X\}^{-1} \{X_0^{\mathrm{T}} X_0 \hat{\beta}_0 + X^{\mathrm{T}} W_R y\},\end{aligned} \tag{13}$$

where W_R is the $n \times n$ diagonal matrix of robust weights.

6.2 Comparison of methods on the bank profit data

Observations with small robust weights are outliers. Agostinelli (2001) suggests a threshold value of 0.5. For the bank profit data, we find 46 observations with weights below 0.5, all of which are also found by the forward search. In the Bayesian analysis using (13), we use the same prior as in the forward search and obtain parameter estimates differing (apart from the last two variables) by no more than 1.3%. The maximum difference is 17%.

The agreement between the two methods is not surprising in this example, where virtually the same set of outliers is declared and the same prior distribution is used. In other examples, such as the Boston housing data (Anglin and Gençay 1996), the differences between the two analyses are greater than those for the bank profit data, but not sufficient to change any conclusions drawn from the analysis of the data. Amongst the comparisons of several methods for frequentist robust regression presented by Riani et al. (2014a), we prefer the forward search because it adds to parameter estimation the monitoring of inferential quantities during the search. As an example, Fig. 6 shows the effect of the outliers which enter towards the end of the search on the HPD regions for the parameters.

7 Power of Bayesian and frequentist procedures

The incorporation of correct prior information into the analysis of data leads to parameter estimates with higher precision than those based just on the sample. There is a consequential increase in the power of tests about the values of the parameters and in the detection of outliers. This section focuses on tests for outlier detection in the presence of correctly and incorrectly specified priors.

We simulate normally distributed observations from a regression model with four variables and a constant ($p = 5$), the values of the explanatory variables having independent standard normal distributions. The simulation envelopes for the distribution of the residuals are invariant to the numerical values of β and σ^2, so we take $\beta_0 = 0$ and $\sigma_0^2 = 1$. The outliers were generated by adding a constant, in the range 0.5 to seven, to a specified proportion of observations, and n_0 was taken as 500. To increase the power of our comparisons, the explanatory variables were generated once for each simulation study. We calculated several measures of power, all of which gave a similar pattern. Here we present results from 10,000 simulations on the average power, that is the average proportion of contaminated observations correctly identified.

Figure 7 shows power curves for Bayesian and frequentist procedures and also for Bayesian procedures with incorrectly specified priors when the contamination rate is 5%. The curves do not cross for powers a little <0.2 and above. The procedure

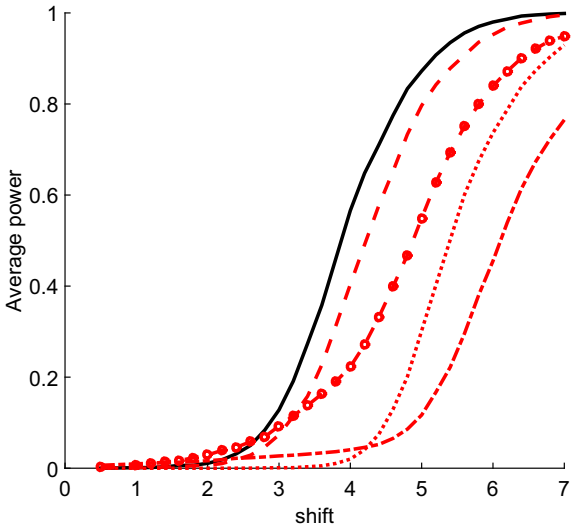

Fig. 7 Average power in the presence and absence of prior information: $\sigma^2 = 1$. Reading across at a power of 0.6: Bayesian, *solid line*; frequentist, *dashed line*; wrong $\beta_0 = -1.5$, *dashed line with circles*; wrong $\sigma_0^2 = 3$, *dotted line*; wrong $\beta_0 = 1.5$, *dotted and dashed line*. Contamination 5%, 2000 simulations, strong prior information; $n_0 = 500$

with highest power is the curve that is furthest to the left which, in the figure, is the correctly specified Bayesian procedure. The next best is the frequentist one, ignoring prior information. The central power curve is that in which the mean of β_0 is wrongly specified as -1.5. This is the most powerful procedure for small shifts, as the incorrect prior is in the opposite direction to the positive quantity used to generate outliers. With large shifts, this effect becomes less important. For most values of average power, the curve for mis-specified σ^2 comes next, with positive mis-specification of β worst. Over these values, three of the four best procedures have power curves which are virtually translated horizontally. However, the curve for mis-specified β has a rather different shape at the lower end caused by the shape of the forward envelopes for minimum deletion residuals. With β mis-specified, the envelopes for large m sometimes lie slightly above the frequentist envelopes. The effect is to give occasional indication of outliers for relatively small values of the shift generating the outliers.

In Fig. 8, for 30% contamination, the Bayesian procedure is appreciably more powerful than the frequentist one, which is slightly less powerful than that with mis-specified σ_0^2. The rule for mis-specified $\beta_0 = 1.5$ has the lowest power, appreciably less than that in which $\beta_0 = -1.5$. Although the curves cross over for shifts around 3.5, the Bayesian procedure with correctly specified prior has the best performance until the shift is sufficiently small that the power is negligible.

8 Discussion

Data do contain outliers. Our Bayesian analysis of the bank profit data has revealed 46 outliers out of 1906 observations. Working backwards from a full fit using single or multiple deletion statistics cannot be relied upon to detect such outliers. Robust methods are essential.

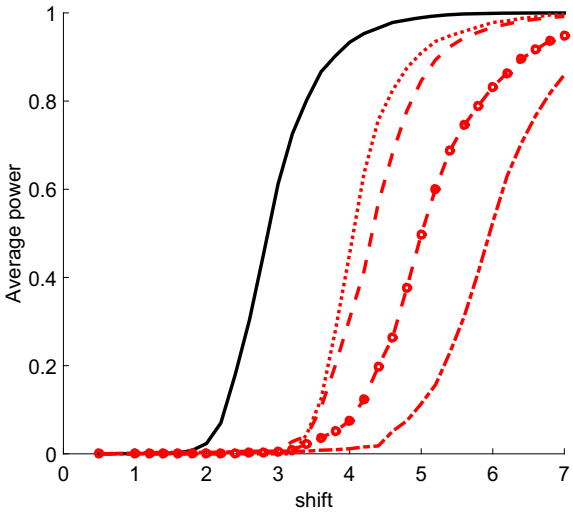

Fig. 8 Average power in the presence and absence of prior information: $\sigma^2 = 1$. Reading across at a power of 0.6: Bayesian, *solid line*; wrong $\sigma_0^2 = 3$, *dotted line*; frequentist, *dashed line*; wrong $\beta_0 = -1.5$, *dashed line with circles*; wrong $\beta_0 = 1.5$, *dotted and dashed line*. Contamination 30%, 2000 simulations, strong prior information; $n_0 = 500$

The results of Sect. 6 indicate how prior information may be introduced into a wide class of methods for robust regression. However, in this paper we have used the forward search as the method of robust regression into which to introduce prior information. There were two main reasons for this choice. One is that our comparisons with other methods of robust regression showed the superiority of the frequentist forward search in terms of power of outlier detection and the closeness of empirical power to the nominal value. A minor advantage is the absence of adjustable parameters; it is not necessary to choose trimming proportion or breakdown point a priori. A second, and very important, advantage is that the structure of the search makes clear the relationship between individual observations entering the search and changes in inferences. This is illustrated in the final part of the plots of parameter estimates and HPD regions in Fig. 6. The structure can also make evident divergencies between prior estimates and the data in the initial part of the search.

A closely-related second application of the method of fictitious observations combined with the FS would be to multivariate analysis. Atkinson et al. (2018) use the frequentist FS for outlier detection and clustering of normally distributed data. The extension to the inclusion of prior information can be expected to bring the advantages of stability and inferential clarity we have seen here.

The advantage of prior information in stabilising inference in the bank profit data is impressive; as we record, the frequentist analysis found 586 outliers. Since many forms of data, for example the bank data, become available annually, statistical value is certainly added by carrying forward, from year to year, the prior information found from previous robust analyses.

Routines for the robust Bayesian regression described here are included in the FSDA toolbox downloadable from http://fsda.jrc.ec.europa.eu/ or http://www.riani.it/ MATLAB. Computation for our analysis of the bank profit data took <10 s on a standard laptop computer. Since, from the expressions for parameter estimation and inference in Sect. 3, the order of complexity of calculation is the same as that for the

frequentist forward search, guidelines for computational time can be taken from Riani et al. (2015).

References

Agostinelli C (2001) `wle`: A package for robust statistics using weighted likelihood. R News 1(3):32–38
Agostinelli C, Greco L (2013) A weighted strategy to handle likelihood uncertainy in Bayesain inference. Comput Stat 28:319–339
Anglin P, Gençay R (1996) Semiparametric estimation of a hedonic price function. J Appl Econ 11:633–648
Atkinson AC (1985) Plots, transformations, and regression. Oxford University Press, Oxford
Atkinson AC, Riani M (2000) Robust diagnostic regression analysis. Springer, New York
Atkinson AC, Riani M, Cerioli A (2010) The forward search: theory and data analysis (with discussion). J Korean Stat Soc 39:117–134. doi:10.1016/j.jkss.2010.02.007
Atkinson AC, Riani M, Cerioli A (2018) Cluster detection and clustering with random start forward searches. J Appl Stat (In press). doi:10.1080/02664763.2017.1310806
Chaloner K, Brant R (1988) A Bayesian approach to outlier detection and residual analysis. Biometrika 75:651–659
Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman and Hall, London
Croux C, Haesbroeck G, Ruwet C (2013) Robust estimation for ordinal regression. J Stat Plan Inference 143:1486–1499
Hampel F, Ronchetti EM, Rousseeuw P, Stahel WA (1986) Robust statistics. Wiley, New York
Hampel FR (1975) Beyond location parameters: robust concepts and methods. Bull Int Stat Inst 46:375–382
Hoffmann I, Serneels S, Filzmoser P, Croux C (2015) Sparse partial robust M regression. Chemom Intell Lab Syst 149(Part A):50–59
Huber PJ, Ronchetti EM (2009) Robust statistics, 2nd edn. Wiley, New York
Johansen S, Nielsen B (2016) Analysis of the forward search using some new results for martingales and empirical processes. Bernoulli 21:1131–1183
Maronna RA, Martin RD, Yohai VJ (2006) Robust statistics: theory and methods. Wiley, Chichester
Pison G, Van Aelst S, Willems G (2002) Small sample corrections for LTS and MCD. Metrika 55:111–123. doi:10.1007/s001840200191
Rao CR (1973) Linear statistical inference and its applications, 2nd edn. Wiley, New York
Riani M, Atkinson AC, Cerioli A (2009) Finding an unknown number of multivariate outliers. J R Stat Soc Ser B 71:447–466
Riani M, Cerioli A, Atkinson AC, Perrotta D (2014a) Monitoring robust regression. Electron J Stat 8:642–673
Riani M, Cerioli A, Torti F (2014b) On consistency factors and efficiency of robust S-estimators. TEST 23:356–387
Riani M, Atkinson AC, Perrotta D (2014c) A parametric framework for the comparison of methods of very robust regression. Stat Sci 29:128–143
Riani M, Perrotta D, Cerioli A (2015) The forward search for very large datasets. J Stat Softw 67(1):1–20
Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79:871–880
Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York
Tallis GM (1963) Elliptical and radial truncation in normal samples. Ann Math Stat 34:940–944
West M (1984) Outlier models and prior distributions in Bayesian linear regression. J R Stat Soc Ser B 46:431–439