136

Reihe Ökonomie

Economics Series

Multivariate Regression and ANOVA Models with Outliers:

A Comparative Approach

Wolfgang Polasek

136

Reihe Ökonomie

Economics Series

Multivariate Regression and ANOVA Models with Outliers:
A Comparative Approach

Wolfgang Polasek

September 2003

Institut für Höhere Studien (IHS), Wien
Institute for Advanced Studies, Vienna

Contact:

Wolfgang Polasek
Department of Economics and Finance
Institute for Advanced Studies
Stumpergasse 56
1060 Vienna, Austria
☎: +43/1/599 91-155
fax: +43/1/599 91-163
email: polasek@ihs.ac.at

Abstract

Assuming a normal-Wishart modelling framework we compare two methods for finding outliers in a multivariate regression (MR) system. One method is the add-1-dummy approach which needs fewer parameters and a model choice criterion while the other method estimates the outlier probability for each observation by a Bernoulli mixing outlier location shift model. For the simple add-1-dummy model the Bayes factors and the posterior probabilities can be calculated explicitly. In the probabilistic mixing model we show how the posterior distribution can be obtained by a Gibbs sampling algorithm. The number of outliers is determined using the marginal likelihood criterion. The methods are compared for test scores of language examination data of Fuller (1987): The results are similar but differ in their strength of their empirical evidence.

Keywords

Multivariate regression, multivariate one-way ANOVA, outliers, Gibbs sampling, marginal likelihoods, sensitivity analysis

JEL Classifications

C11, C39

Comments

The author wants to thank two unknown referees for their comments. The program can be found in the program page BASEL under http://www.unibas.ch/iso/basel.

Contents

1 Introduction

Multivariate regression (MR) models are a popular tool in social sciences (e.g., economics, psychology and sociology) for explaining a multivariate data matrix by a common set of "independent" variables or regressors. Such an approach is particularly useful if one encounters many variables (repeated measures) which can be related by some common variables. The literature on multivariate outliers is surprisingly short, and in this paper we demonstrate a computational Bayesian approach.

In particular we are interested as to whether or not the new technique of Monte Carlo Markov Chain (MCMC) methods can be used to detect multivariate outliers and if they can beat simpler models of outlier detections in a multivariate context. Using the Gibbs sampling approach of Verdinelli and Wasserman (1991) for multivariate location shift outlier models, we show how to derive the marginal likelihoods from the Gibbs sampling algorithm.

Marginal likelihoods are used for Bayesian testing and are a widely used tool for model selection. Therefore, they can be used to determine the number of outliers in a univariate or multivariate regression (MR) model. The ratio of marginal likelihoods defines the Bayes factor and for simple models this ratio can be obtained in closed form. For the more complicated probabilistic outlier model we will show how the approach of Chib (1995) can be used to calculate marginal likelihoods from the MCMC simulation output. The probabilistic outlier model is a mixture model for each observation in the sample: The ordinary regression model is contaminated by an location shift outlier model, where the mixing probability follows a Bernoulli distribution with unknown parameter. In the Bernoulli mixing location shift outlier model, which we will call briefly the "probabilistic MR outlier model" this approach is possible, since all the full conditional distributions of the MR model can be derived in closed form.

Though various approaches to outlier modelling can be found in the literature (see, e.g., Kitagawa and Akaike (1982), Barnett and Lewis (1984), Pettit and Smith (1985)), little work has been done for outlier models in multivariate Bayesian analysis. Therefore we will use the simple location shift outlier model as a basic model for detecting outliers in regression analysis. In univariate comparisons, location shift outlier models have been found superior to, e.g., multiplicative or variance inflation outliers because they produce more likely aberrant observations than "inliers". In order to constrain the computational burden, a simple probabilistic outlier model is needed to implement the Gibbs sampler but also to facilitate the computational burden for model choice.

The plan of the paper is as follows: In section two we introduce the probabilistic Bayesian regression model with additive outliers. This is a multivariate extension of the outlier approach of Verdinelli and Wasserman (1991). The marginal likelihoods needed for model testing and the number of factors are derived in section three. This also allows to determine the number of regressor variables in a multivariate regression model. In section four we analyze the language data set of Fuller (1987) and compare inferences for possible (location shift) outliers.

Section five concludes and in the appendix we have listed an explicit result as how to obtain the marginal likelihood in an informative multivariate regression model (based on results of Polasek and Ren (1998)). Thus, the marginal likelihood of the MR+outlier model (or add-1-dummy variable model) is used as a simple benchmark for outlier modelling, and additionally, will be compared with the probabilistic MR outlier modelling approach.

In summary, we will show that marginal likelihoods can be used as a powerful criterion for modelling outliers: Because it can be calculated for any model and any parametrization it can used not only for finding outliers but also to answer which model is the best and how many outliers are possibly present. Like information criteria the marginal likelihoods tend to pick the parsimonious model and allows also to quantify the presence of outliers by posterior probabilities.

2 Multivariate regression (MR) analysis with outliers

Consider the MR analysis model, i.e. n observations consisting of a row response vector $\mathbf{y}_i = (y_1, ..., y_p) : 1 \times p$ of length p where each response variable is explained by a K-dimensional regressor $\mathbf{x}_i = (x_1, ..., x_K) : 1 \times K$ and we assume a multivariate normal distribution for the error term

$$\mathbf{y}_i \sim \mathcal{N}_p[\mathbf{x}_i\mathbf{B}, \mathbf{\Psi}], \quad i = 1, \ldots, n,$$

where $\mathbf{B}$ is a $K \times p$ matrix of regression coefficients and $\mathbf{\Psi}$ is a $p \times p$ symmetric covariance matrix of the observations. We now explain how this model is extended to the "probabilistic MR outlier model", i.e. more precisely the Bernoulli mixing location shift outlier model. The first model of this sort was the univariate location shift outlier model that was analyzed by the Gibbs sampler in Verdinelli and Wasserman (1991). The multivariate location shift outlier model is formulated with the n Bernoulli distributed indicator variables $\vartheta_1, \ldots, \vartheta_n$

$$\begin{aligned} f(\mathbf{y}_i) &= P(\vartheta_i = 0) f(\mathbf{y}_i \mid \vartheta_i = 0) + P(\vartheta_i = 1) f(\mathbf{y}_i \mid \vartheta_i = 1) & (1) \\ &= (1 - \varepsilon_*)\, \mathcal{N}_p\, [\mathbf{y}_i \mid \mathbf{x}_i \mathbf{B}, \mathbf{\Psi}] + \varepsilon_*\, \mathcal{N}_p\, [\mathbf{y}_i \mid \mathbf{a}_i + \mathbf{x}_i \mathbf{B}, \mathbf{\Psi}] & (2) \end{aligned}$$

where $\mathbf{y}_i$ is the i-th row of the $n \times p$ observation matrix $\mathbf{Y}$ with $\mathbf{Y}' = (\mathbf{y}_1', \ldots, \mathbf{y}_n')$, and $\mathbf{a}_i : 1 \times p$ is the i-th row of the $n \times p$ location shift matrix $\mathbf{A}$ with $\mathbf{A}' = (\mathbf{a}_1', \ldots, \mathbf{a}_n')$. $\mathbf{D}_\vartheta = diag(\vartheta_1, \ldots, \vartheta_n)$ is a $n \times n$ indicator matrix for the (multivariate) outliers. We assume that each indicator is distributed as a Bernoulli random variable with parameter ε_*, the prior (mixing or "probability") parameter that the i-th observation is an outlier.
Assuming independence between the observations, the probabilistic MR model with outliers is given by

$$\mathbf{Y} \;\sim\; \mathcal{N}_{n \times p}[\mathbf{XB} + \mathbf{D}_\vartheta \mathbf{A}, \mathbf{\Psi} \otimes \mathbf{I}_n] \qquad (3)$$

and can be written in transposed form as

$$\mathbf{Y}' \;\sim\; \mathcal{N}_{p \times n}[\mathbf{B}'\mathbf{X}' + \mathbf{A}'\mathbf{D}_\vartheta, \mathbf{I}_n \otimes \mathbf{\Psi}] \qquad (4)$$

where the $n \times K$ regressor matrix $\mathbf{X}$ is defined as $\mathbf{X}' = (\mathbf{x}_1', \ldots, \mathbf{x}_n')$. The prior information for the parameter matrices can be compactly formulated as

$$\begin{aligned} \mathbf{B} &\sim \mathcal{N}_{K \times p}[\mathbf{B}_*, \mathbf{G}_* \otimes \mathbf{H}_*], \\ \mathbf{\Psi}^{-1} &\sim \mathcal{W}_p[\mathbf{\Psi}_*, n_*], \\ \mathbf{A} &\sim \mathcal{N}_{n \times p}[\mathbf{A}_*, \mathbf{P}_* \otimes \mathbf{I}_n], \\ \vartheta_i &\sim Ber[\varepsilon_{i*}], \qquad i = 1, \ldots, n, \end{aligned}$$

where $Ber[\varepsilon_{i*}]$ denotes the Bernoulli distribution and ε_{i*} is the prior probability parameter that observation i is an outlier.
Furthermore, $\mathbf{A}_* : n \times p$ and $\mathbf{P}_* : p \times p$ are a-priori known parameter matrices for the location and variances of outliers and ε_{i*} is the known "success" probability of being an outlier, mostly set to a constant value ε_*.

2.1 Gibbs sampling in the probabilistic MR+outlier model

The multivariate regression model is given by

$$\underset{(n\times p)}{\mathbf{Y}} = \underset{(n\times K)}{\mathbf{X}} \; \underset{(K\times p)}{\mathbf{B}} + \underset{(n\times p)}{\mathbf{E}}.$$

Let $\mathbf{A}$ be a $n \times p$ location shift parameter matrix and $\mathbf{D}_\vartheta = diag(\vartheta_1, \ldots, \vartheta_n)$ an $n \times n$ indicator matrix for multivariate outliers. We can then formulate the model by assuming a normal distribution

$$\mathbf{Y} \sim \mathcal{N}_{n\times p}[\mathbf{XB} + \mathbf{D}_\vartheta \mathbf{A}, \mathbf{\Psi} \otimes \mathbf{I}_n].$$

The joint distribution of the data $\mathbf{Y}$ and the parameter $\theta = (\mathbf{B}, \mathbf{\Psi}, \mathbf{A}, \mathbf{D}_\vartheta)$ is

$$\begin{aligned} p(\mathbf{Y}, \theta) &= \mathcal{N}_{n\times p}[\mathbf{Y} \mid \mathbf{XB} + \mathbf{D}_\vartheta \mathbf{A}, \mathbf{\Psi} \otimes \mathbf{I}_n] \cdot \mathcal{N}_{\mathbf{K\times p}}[\mathbf{B}|\mathbf{B}_*, \mathbf{G}_* \otimes \mathbf{H}_*] \cdot \\ & \cdot \mathcal{W}_p[\mathbf{\Psi}^{-1} \mid \mathbf{\Psi}_*, n_*] \cdot \mathcal{N}_{n\times p}[\mathbf{A} \mid \mathbf{A}_*, \mathbf{P}_* \otimes \mathbf{I}_n] \cdot \sum_{i=1}^{n} Ber(\vartheta_i \mid \varepsilon_{i*}). \end{aligned}$$

We use θ^c as conditional argument in a full conditional distribution to denote the parameter set θ without the current parameter argument. The full conditional distributions for the Gibbs sampler are

a) For the matrix regression coefficients $\mathbf{B}$:

$$p(\mathbf{B} \mid \mathbf{Y}, \theta^c) = \mathcal{N}_{K\times p}[\mathbf{B}_{**}, \mathbf{C}_{**}]$$

a multivariate normal distribution with the parameters

$$\begin{aligned} \mathbf{C}_{**}^{-1} &= \mathbf{G}_*^{-1} \otimes \mathbf{H}_*^{-1} + \mathbf{\Psi}^{-1} \otimes \mathbf{X}'\mathbf{X}, \\ \text{vec } \mathbf{B}_{**} &= \mathbf{C}_{**}[\text{vec } (\mathbf{G}_*^{-1}\mathbf{B}_*\mathbf{H}_*^{-1} + \mathbf{X}'(\mathbf{Y} - \mathbf{D}_\vartheta \mathbf{A})\mathbf{\Psi}^{-1})]. \end{aligned}$$

b) For the covariance matrix $\mathbf{\Psi}$:

$$p(\mathbf{\Psi}^{-1} \mid \mathbf{Y}, \theta^c) = \mathcal{W}_p[\mathbf{\Psi}_{**}, n_{**} = n_* + n]$$

a p-dimensional Wishart distribution with scale parameter

$$\mathbf{\Psi}_{**} = \mathbf{\Psi}_* + (\mathbf{Y} - \mathbf{XB} - \mathbf{D}_\vartheta \mathbf{A})'(\mathbf{Y} - \mathbf{XB} - \mathbf{D}_\vartheta \mathbf{A}).$$

c) For the level shift matrix $\mathbf{A}$:

$$p(\mathbf{A} \mid \mathbf{Y}, \theta^c) = \mathcal{N}_{n\times p}[\mathbf{A}_{**}, \mathbf{G}_{**}]$$

a multivariate normal distribution with the parameters

$$\begin{aligned} \mathbf{G}_{**}^{-1} &= \mathbf{P}_*^{-1} \otimes \mathbf{I}_n + \mathbf{\Psi}^{-1} \otimes \mathbf{D}'_\vartheta \mathbf{D}_\vartheta, \\ \text{vec } \mathbf{A}_{**} &= \mathbf{G}_{**}[\text{vec } (\mathbf{A}_* \mathbf{P}_*^{-1} + \mathbf{D}_\vartheta(\mathbf{Y} - \mathbf{XB})\mathbf{\Psi}^{-1})]. \end{aligned}$$

For each observation the posterior mean can be calculated by breaking up the system into univariate equations:

$$\mathbf{G}_{**i}^{-1} = \mathbf{P}_*^{-1} + {\vartheta_i}^2 \mathbf{\Psi}^{-1}, \quad i = 1, \ldots, n,$$

$$\mathbf{a}_{**i} = \mathbf{G}_{**i}[\mathbf{P}_*^{-1}\mathbf{a}_{*i} + \vartheta_i \mathbf{\Psi}^{-1}(\mathbf{y}_i - \mathbf{B}\mathbf{x}_i)].$$

d) For the indicator variables ϑ_i:

$$Pr(\vartheta_i \mid \mathbf{Y}, \theta^c) = Ber[\varepsilon_{i**} = \frac{c_i}{c_i + d_i}], \qquad i = 1, \ldots, n,$$

a Bernoulli distribution with the components obtained via Bayes theorem, i.e.,

$$\begin{aligned} c_i &= \mathcal{N}_p[\mathbf{y}_i \mid \mathbf{x}_i \mathbf{B} + \mathbf{a}_i, \mathbf{\Psi}] \cdot \varepsilon_{i*}, \\ d_i &= \mathcal{N}_p[\mathbf{y}_i \mid \mathbf{x}_i \mathbf{B}, \mathbf{\Psi}] \cdot (1 - \varepsilon_{i*}), \qquad i = 1, \ldots, n, \end{aligned}$$

where $\mathbf{x}_i$ is the i-th row of $\mathbf{X}$ and $\mathbf{a}_i$ is the i-th row of $\mathbf{A}$.

3 The marginal likelihood for the MR+outlier model

Using the approach of Chib (1995) we will evaluate the marginal likelihood at the point

$$\hat{\theta}_1 = (\hat{\theta}_0, \hat{\mathbf{A}} = \mathbf{0}, \hat{\mathbf{D}}_\vartheta = \mathbf{0}) \tag{5}$$

where $\hat{\theta}_0 = (\hat{\mathbf{B}}, \hat{\boldsymbol{\Psi}})$ is the same point as for the MR without outliers. Therefore we have the following factorization

$$p(\hat{\theta} \mid \mathbf{Y}) \;=\; p(\hat{\mathbf{D}}_\vartheta \mid \mathbf{Y}) \cdot p(\hat{\mathbf{A}} \mid \hat{\mathbf{D}}_\vartheta, \mathbf{Y}) \cdot p(\hat{\theta}_0 \mid \hat{\mathbf{A}}, \hat{\mathbf{D}}_\vartheta, \mathbf{Y}) \tag{6}$$

and

$$p(\hat{\theta}) \;=\; p(\hat{\theta}_0)\mathcal{N}[\mathbf{A}_*, \mathbf{I}_n \otimes \mathbf{P}_*] \prod_{i=1}^{n} Ber(\varepsilon_*). \tag{7}$$

1. Use the Gibbs run of J sample points of the 'MR+outlier' program to calculate the ordinate:

$$\begin{aligned} p(\hat{\mathbf{D}}_\vartheta \mid \mathbf{Y}) &= \int \prod_{i=1}^{n} Ber(\varepsilon_{i**}) p(\theta \mid \mathbf{Y}) d\theta \\ &= \frac{1}{J} \sum_{j=1}^{J} \prod_{i=1}^{n} Ber(\varepsilon_{i**}^{(j)}) \end{aligned} \tag{8}$$

where the parameter of the i-th posterior density of the Bernoulli distribution is given by

$$\begin{aligned} \varepsilon_{i**}^{(j)} &= \frac{c_i^{(j)}}{c_i^{(j)} + d_i^{(j)}}, \\ \textit{with} \quad c_i^{(j)} &= \mathcal{N}_p[\mathbf{y}_i \mid \mathbf{a}^{(j)} \vartheta_i^{(j)} + \boldsymbol{\Lambda}_j \mathbf{z}_i^{(j)}], \\ \\ d_i^{(j)} &= \mathcal{N}_p[\mathbf{y}_i \mid \boldsymbol{\Lambda}^{(j)} \mathbf{z}_i^{(j)}]. \end{aligned}$$

2. The ordinate for the second component can be obtained without a Gibbs sampling output by

$$p(\hat{\mathbf{A}} \mid \hat{\mathbf{D}}_\vartheta = \mathbf{0}, \mathbf{Y}) = \mathcal{N}_{n \times p}[\hat{\mathbf{A}} \mid \mathbf{A}_{**}, \mathbf{I}_n \otimes \mathbf{G}_{**}] = \prod_{i=1}^{n} \mathcal{N}[\hat{\mathbf{a}}_i | \mathbf{a}_{i**}, \mathbf{G}_{**}]. \tag{9}$$

It can be seen from the full conditional distribution for the location shift parameter $\mathbf{A}$ that setting $\hat{\mathbf{D}}_\vartheta = \mathbf{0}$ the conditional distribution equals the prior distribution:

$$\mathbf{G}_{**} = \mathbf{I}_n \otimes \mathbf{P}_* \qquad \text{and} \qquad vec\ \mathbf{A}_{**} = vec\ \mathbf{A}_*.$$

3. Finally we can obtain the ordinate of the third factor $p(\hat{\theta}_0 \mid \hat{\mathbf{A}}, \hat{\mathbf{D}}_\vartheta, \mathbf{Y})$ in (6) by the marginal likelihood calculations of a MR model without outliers. In appendix A it is shown how this marginal likelihood can be calculated in explicit form. (For further possibilities on model selection by marginal likelihoods see Polasek and Ren (1998).
Calculations become simpler if we use the marginal likelihoods in logarithmic form

$$\log p(\mathbf{Y}) = log\ p(\mathbf{Y} \mid \hat{\theta}_1) + log\ p(\hat{\theta}_1) - log\ p(\hat{\theta} \mid \mathbf{Y}) \tag{10}$$

where the likelihood part is given by

$$p(\mathbf{Y} \mid \hat{\theta}_1) = \mathcal{N}[\mathbf{Y} \mid \hat{\mathbf{X}}\hat{\mathbf{B}}, \hat{\mathbf{\Psi}} \otimes \mathbf{I}_n], \tag{11}$$

which is the same value as for the MR without outliers, since $\hat{\mathbf{D}}_\vartheta = \mathbf{0}$.
Note that formula (8) is a simplification since the components for the location shifts $\hat{\mathbf{A}}$ in (10) cancel out.

3.1 Model selection with Bayes factors

Posterior odds are used in Bayesian analysis to choose between two or more different models for the same data set. The basic formula for choosing between models M_1 and M_2 is

$$\begin{array}{rcl} \text{posterior odds} & = & \text{Bayes factor} \cdot \text{prior odds} \\ & \text{or} & \\ \dfrac{p(M_1|\mathbf{Y})}{p(M_2|\mathbf{Y})} & = & B \cdot \dfrac{p(M_1)}{p(M_2)}, \end{array} \tag{12}$$

where $p(M_1|\mathbf{Y})$ and $p(M_2|\mathbf{Y})$ are the posterior probabilities for models M_1 and M_2, respectively. $p(M_1)$ and $p(M_2)$ are the prior probabilities for models M_1 and M_2, and, in the simplest case, they are set to be equal. Thus, in these cases the posterior odds are equal to the Bayes factor, which is defined as the ratio of marginal likelihoods

$$B = \frac{p(M_1|\mathbf{Y})}{p(M_2|\mathbf{Y})} = \frac{\int p(\mathbf{Y}, \theta_1) d\theta_1}{\int p(\mathbf{Y}, \theta_2) d\theta_2}, \tag{13}$$

where θ_1 and θ_2 are the parameters for models M_1 and M_2, respectively. If $B > 1$ we choose model M_1 and if $B < 1$ we choose model M_2. Therefore the model with the largest marginal likelihood will be chosen using simple Bayes factors. For example: The Bayes factor for testing the factor analysis model with outliers against no outliers is:

$$B = \frac{p(\mathbf{Y}|\text{outliers})}{p(\mathbf{Y}|\text{no outliers})}.$$

4 Example

We use the language data in Fuller (1987, page 154) as an example for identifying outliers in a multivariate regression analysis. This data consists of 100 observations with eight variables (measured on a scale which is assumed to be approximately continuous: it is the sum of marks on 2 essays, i.e. the range is 1-10): the first three variables related to the way the essay is written (poorly developed - well developed, difficult to understand - easy to understand, illogical-logical), three variables related to the way how the language used (inappropriate - appropriate, unacceptable - acceptable, irritating - not irritating), and two variables related to the writing style (careless - careful, unintelligent - intelligent).

We are interested to see if we can detect outliers in this data set of typical school subject scores. The multivariate approach seem to be particularly interesting since the scores of one person of an essay by 8 categories might be not independent. Knowing if a particular data set (like this 8-dimensional variable set) contains outliers might be important for any further analysis (e.g. like factor analysis or cluster analysis).

To make the multivariate model as simple as possible, we have used as independent regressor variables only the intercept which reduces the multivariate regression model to a multivariate one-way ANOVA model.

The prior distribution for the MR model (1) consists of two parts. The first part is the set of parameters which is identical to the MR model without outliers and in the second part we elicit the parameters for the outlier model. The prior mean is simply set to the mid point of the valuation scale which is 5. Therefore the prior mean matrix reduces to a vector, i.e. $\mathbf{B}_* = 5 * \mathbf{1}_k$. Furthermore, we assume that the residual variances of the factor model are about one tenth of the variances of the observed variables. The value of the prior information of the Wishart distribution is $n_* = \nu_* = 1$, i.e. 1/100 in terms of the sample size $n = 100$.

$$\mathbf{\Psi}_* \; = \; diag(var(\mathbf{y}_1), \ldots, var(\mathbf{y}_p))/10.$$

For the prior distribution of the location shifts we have assumed $\mathbf{A}_* = \mathbf{0}$ and $var(\mathbf{a}_i) = \mathbf{P}_* = diag(var(\mathbf{y}))$. Convergence of the Gibbs sampler was achieved quite quickly for the present specification. The convergence was monitored by diagnostic measures proposed in the CODA package of Best et al. (1995) written in S-plus which uses the Gelman and Rubin (1992) and the Raftery and Lewis (1992) statistics. A good introduction to the theory and practice of MCMC modelling can be found in Gilks et al. (1990). Only the last 100 simulations of the MCMC sequence were used to calculate the mean and variances of the posterior distribution.

4.1 The results of the add-1-dummy MR outlier model

Simple outlier detection is possible by adding a dummy variable to the regressor matrix of a regression model. When the 1-location of the dummy varies over all possible observations we obtain n different model estimates and we are confronted with a model choice problem. Assuming equal probabilities for all these Bayesian regression models we can use the marginal likelihoods as model choice criterion.

Table 1 calculates all the marginal likelihoods with the MR outlier model based on the results of the appendix A (Polasek and Ren (1998)) which extends the MR model by a single dummy variable. If the marginal likelihood is calculated successively for all observations (from 1 to 100) then we obtain a plot of marginal likelihoods as in Figure 1. The peaks of this plot show the observation numbers which are potential outliers.

Comparing the log-marginal likelihoods by a ratio leads to a Bayes test. Bayes factors (BFs) can be judged by the $9:19:99$ rule, assuming that the more likely model is in the nominator: A $BF > 9$ is remarkable, $BF > 19$ is significant and a $BF > 99$ is highly significant. Transforming this rule to the log scale, we just have to calculate the differences between marginal likelihoods:

$$lnBF = lnf(\mathbf{Y}|model_1) - lnf(\mathbf{Y}|model_2).$$

On the log scale the cut-off points of the $9:19:99$ rule are $ln9 = 2.2, ln19 = 2.9$ and $ln99 = 4.6$ or in short: $2, 3$ and 5.

Model	obs.	log. marginal likelihood
MR	no	-1317.0456
MR-outlier	8	-1306.8017
	11	-1299.397
	31	-1305.674
	37	-1304.267
	80	-1304.086
	85	-1307.717

Table 1: The log.-marginal likelihood of the language data in Fuller (1987) for the MR and 6 MR-outlier models.

For a formal Bayes test we just have to compare the log-marginal likelihoods in Table 1. The MR model with no outliers (the "null model") has a log-marginal likelihoods of -1317.0. The MR model with observation #8 as an outlier has a log-marginal likelihood of -1306.8. The difference is 10.2 and we infer from the log cut-off points that this result is highly significant (given a prior probability that both models are equally likely). In the same way we do a pairwise comparison of all the 6 potential outlier locations and compare them with the MR model without outliers. All outliers locations are highly significant and the highest significance is found for observation number 11. (Note that it would be also possible to test all the observations jointly for being outliers. This requires the extension of the MR model by 6 dummy variables and is not the main goal of this analysis.)

4.2 The results of the probabilistic MR outlier model

Figure 3 plots all the posterior probabilities that a certain observation is an outlier from the probabilistic Gibbs sampling model. All the observations for which the probabilities are greater than .5 can be found in column 2 of Table 2. It can be seen that the set of observations which is classified as outliers by the two procedures, is the same. Only the implicit evaluation what is significant is quite different between these 2 approaches. Since the Gibbs sampling approach is based on a simultaneous estimation model with

Obs.	Prob.	Location Shift (std)			
		Developed	Logical	Irritating	Intelligent
8	0.9189	**0.5157(1.3461)**	-0.1012(0.8603)	-0.0752(1.5642)	-0.1201(0.9854)
11	0.5522	-0.2088(0.9334)	0.1395(1.0421)	-0.3051(1.0641)	**0.1337(0.0998)**
31	0.5477	-0.0572(0.9214)	-0.2612(1.3041)	-0.5623(1.4952)	0.3357(1.0911)
37	0.6258	-0.1265(1.1921)	-0.1919(1.2320)	-0.2093(1.4961)	0.1751(1.3272)
80	0.5227	-0.0183(1.3321)	0.0591(0.9468)	0.0820(0.7019)	0.0844(0.7635)
85	0.5153	0.6850(1.2030)	-0.1029(0.7405)	0.0555(1.1533)	-0.1935(1.2412)

Obs.	Prob.	Location Shift (std)			
		Understand	Appropriate	Acceptable	Careful
8	0.9189	-0.1109(1.1913)	-0.0193(0.7842)	-0.1090(1.2711)	-0.1177(1.0913)
11	0.5522	0.4183(0.8916)	-0.1615(1.4734)	0.0185(1.3451)	0.0725(1.3422)
31	0.5477	0.4740(1.15512)	-0.1983(1.0581)	**0.2741(1.4472)**	0.0166(1.4311)
37	0.6258	-0.0116(1.411)	-0.1313(1.0091)	**0.1389(0.1296)**	**-0.1983(0.1122)**
80	0.5227	-0.1012(1.2631)	**0.1119(0.1061)**	-0.2566(1.0832)	0.0407(0.8509)
85	0.5153	-0.1753(1.3020)	0.0102(1.4132)	0.3242(0.8961)	-0.0278(1.1960)

Table 2: The probability of being an outlier and the posterior mean of location shifts and standard deviations in MR with outliers model of the language data in Fuller (1987). Posterior means larger than posterior standard deviations are bold face.

many more parameters, it is not surprising that the "significant" results are less occurring.

Table 2 shows the rows of the estimated location shift matrix $\mathbf{A}$ (precisely it is the posterior mean) for which the posterior probability parameter ε_{i**} (the probability of being an outlier) is larger than 1/2. The prior probability that observation i is an outlier is assumed to be $\varepsilon_{i*} = 0.1$. The (posterior) standard deviations of the location shifts are printed in parentheses. Those location shifts a_{ij} which are larger than the standard deviation are printed in bold font. It is interesting to note that all five outlier points have location shifts which are shifted by more than one standard deviation in exactly one of the eight variables. This shows that the grading process of the language papers was quite independent with respect to these eight judgment variables. No outlier point shows remarkable or significant location shifts in *two* or more variables jointly. Note that the standard deviations of the location shifts varies quite a lot across the outliers. There seems to be no obvious relation to the posterior probability of being an outlier. The size of the location shifts are generally not too large and lie in a reasonable range when compared to the original data.

Figure 2 shows the posterior distribution of the location shifts by parallel box plots. The posterior means ε_{i**} are plotted in Figure 3 and are interpreted as posterior probabilities of being an outlier. This leads us to the following stochastic outlier analysis: While the posterior means for three observations are above 60%, two more observations just lie above the 50% line. Three (or almost four) additional observations are above 40% while the other observations are certainly not candidates for outlier locations. We conclude that checking for outliers can be important for multivariate regression when there are data sets that contain possibly aberrant observations.

4.3 Prior probabilities and sensitivity analysis

Sensitivity analysis is important in all those cases where prior probabilities might have a large influence on the results. Since the outlier model with the Gibbs sampler estimates a model with a outlier parameter for each observation we are faced with a problem with many parameters. Therefore it is not surprising that the Gibbs-outlier model in Table 2 comes up with smaller probabilities that a single observation can be an outlier.
Now we can perform the following thought experiment: Given the posterior probabilities of the Gibbs outlier model and the Bayes factor of the add-1-dummy MR model, what prior probabilities of the ”null model” could have generated these results?
The posterior odds W of an outlier model is obtained from the Bayes factor BF by $W = BF * w$ (in (12) and in logs this relationship is given by

$$logW = logBF + logw$$

where w is the prior odds, i.e.

$$w = P(outlier)/(1 - P(outlier)).$$

Given the posterior odds W from the Gibbs sampling model and the BF from the non-Gibbs model, we can calculate the prior odds which is given by

$$Logw = logW_{Gibbs} - logBF_{non-Gibbs}.$$

From the value of the prior odds w we can calculate $P(outlier)$. The results of this analysis can be seen in the next table.

	Hypothetical outlier probabilities		
Obs.	Gibbs-prob.	log. BF	prior prob.
8	.9189	10.2439	0.00040291
11	.5522	17.6490*	2.6678E-08
31	.5477	11.3718	1.3944E-05
37	.6258	12.9597	3.9355E-06
80	.5227	12.9597	2.5771E-06
85	.5153	9.3284	9.4466E-05

Table 3: Sensitivity analysis with respect to prior probabilities
(The asterisk marks the observation with the highest BF)

The first column shows the observation number of the outlier in the data set (out of 100 observations). The second column reports the posterior probabilities to be an outlier from the Gibbs sampling model. The third column shows the Bayes factor of being an outlier based on the non-Gibbs model and the final column calculates the hypothetical prior probabilities for the MR-outlier model to match the probabilities of the Gibbs model.

Note that if the prior odds w are 1, i.e. it is equally likely that an outlier exists at a certain location or not, then the large BF produces a high probability that the outlier model is correct. If we assign a smaller probability that just this observation could be an outlier (i.e. under equal possibilities a 1/100 chance) then the influence of the BF is weakened. How low this probability has to become to get the Gibbs sampler probability is shown in the last column of Table 3.

As we see from the last column, the prior probabilities that an outlier model exists would have to be really small (or the probability of the null model really large) to obtain the posterior probability results of the Gibbs sampler. Thus we conclude that the prior probabilities are not influencing the empirical evidence for the non-Gibbs MR model: Any reasonable probability of the existence of an outlier model would have given overwhelming evidence that there exist an outlier on this location.

5 Summary

The paper has compared two approaches for outlier detection in a Bayesian multivariate regression model which we have called the probabilistic MR+outlier model and the add-1-dummy MR model. Under the usual assumption of a normal-Wishart distribution and a location-shift outlier model with Bernoulli distributed indicator variables, all the full conditional distributions of the Gibbs sampler can be derived in closed form. A simpler model for outlier detection is the estimation of a dummy variable in the MR model for all possible observations, i.e. the add-1-dummy method. While the Gibbs model estimates the probability of seeing an outlier on the j-th position, the add-1-dummy MR model gives a hypothesis probability in a series of model choice. Both methods have been demonstrated with the Fuller (1987) language data set and the presence of outliers is explored probabilistically. The marginal likelihood can be computed in both approaches and are used for Bayes tests. Both methods yield the same set of outliers but differ in the strength of their empirical evidence. The results show that the model with fewer parameter produces higher posterior probabilities for the presence of outliers. Further research in this area will show if MR models can be analyzed unsuccessfully or more efficiently by different MCMC strategies (which allows faster model comparisons) or different outliers modelling approaches, like different distributional assumptions.

6 References

Barnett, V. and Lewis, T. (1984): *Outliers in Statistical Data*, Wiley, Chichester.

Best, N.G.; Cowles, M.K. and Vines, K. (1995): CODA: Convergence diagnosis and output analysis software for Gibbs sampling output, Version 3.0, Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.

Chib, S. (1995): Marginal likelihood from the Gibbs output. *JASA, 90*, 1313-1321.

Fuller, W.A. (1987): *Measurement Error Models*, John Wiley & Sons, NY.

Gelfand, A.E. and Smith, A.F.M. (1990): Sampling based approaches to calculating marginal densities. *Journal of American Statistical Association, 85*, 398-409.

Gelman, A. and Rubin, D.B. (1992): Inference from iterative simulation using multiple sequences (with discussion), *Stat. Science, 7*, 457-511.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1990): Markov Chain Monte Carlo in Practice, Chapman and Hall, London.

Kitagawa, G. and Akaike, H. (1992): A Quasi Bayesian Approach to Outlier Detection, *Ann. Inst. Stat. Mathematics, 34*, 95-104.

Magnus, J.R. and Neudecker, H. (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley, Chichester.

Pettit, L.I. and Smith, A.F.M. (1985): Outliers and Influential Observations in Linear Models, in J.M. Bernardo et al. (eds.) *Bayesian Statistics, 2*, 473-474, Elsevier Publ. co.

Polasek, W. and Ren L. (1998): Structural breaks and model selection with marginal likelihoods, in: W. Racugno (ed). Proceedings of the Workshop on Model Selection, Pitagora Ed. Bologna, 223-273, with discussion.

Polasek, W. (ed.) (1998): The BASEL package, ISO-WWZ, University of Basel, mimeo.*http:www.ihs.ac.at/ polasek*

Raftery, A.E. and Lewis, S.M. (1992): One long run with diagnostics: implementation strategies for Markov chain Monte Carlo, *Stat. Science, 7*, 493-497.

Verdinelli, I. and Wasserman, L. (1991): Bayesian analysis of outlier problems using the Gibbs sampler, *Statistics and Computing 1991-1*, 105-117.

A Appendix: The marginal likelihood of MR models for informative priors

Theorem A.1 The marginal likelihood for informative priors Consider the regression model for the $n \times p$ matrix Y

$$Y = XB + U, \tag{14}$$

or

$$Y \sim N[XB, \Sigma \otimes V_*],$$

with the regressor matrix $X : (n \times K)$, coefficient matrix $B : (K \times Mp)$ and V_* is a known covariance matrix (e.g. $V_* = I_n$). Assuming a normal-Wishart prior

$$f(B, \Sigma^{-1}) = NW[B_*, H_*, \Sigma_*, n_*],$$

then the marginal likelihood is

$$f(Y) = (2\pi)^{-\frac{np}{2}} \frac{c_{n_{**}}}{c_{n_*}} \frac{|\Sigma_*|^{\frac{n_*}{2}}}{|\Sigma_{**}|^{\frac{n_{**}}{2}}} \frac{|H_{**}|^{\frac{p}{2}}}{|H_*|^{\frac{p}{2}}} \tag{15}$$

with

$$\Sigma_{**} = \Sigma_* + \hat{U}'\hat{U} + \Delta, \quad \hat{U} = Y - X\hat{B},$$

$$H_{**}^{-1} = X'X + H_*^{-1},$$

$$\Delta = (\hat{B} - B_*)'[(X'X)^{-1} + H_*]^{-1}(\hat{B} - B_*),$$

$$\hat{B} = (X'X)^{-1}X'Y,$$

$$n_{**} = n_* + n,$$

and the constant c_{n_*} is given by

$$c_{n_*} = 2^{\frac{pn_*}{2}} \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma[\frac{n_* + 1 - j}{2}]. \tag{16}$$

The constant $c_{n_{**}}$ is given in similar way, where n_{**} replaces n_*. Note that the marginal likelihood does not depend on the number of regressors K.
Proof: The likelihood function is

$$\begin{aligned} f(Y|B,\Sigma^{-1}) &= N[Y|XB, \Sigma \otimes I_n] \\ &= (2\pi)^{-\frac{np}{2}} |\Sigma \otimes I_n|^{-1/2} \exp\{-\frac{1}{2} tr \Sigma^{-1}(Y-XB)'(Y-XB)\}. \end{aligned}$$

The prior distribution is

$$\begin{aligned} f(B,\Sigma^{-1}) &= N[B|B_*, \Sigma \otimes H_*] \quad \Gamma[\Sigma^{-1}|\Sigma_*, n_*] \\ &= (2\pi)^{-\frac{Kp}{2}} |\Sigma \otimes H_*|^{-\frac{1}{2}} \exp\{-\frac{1}{2} tr \Sigma^{-1}(B-B_*)' H_*^{-1}(B-B_*)\} \cdot \\ &\quad \cdot c_{n*}^{-1} \cdot |\Sigma_*|^{\frac{n_*}{2}} |\Sigma^{-1}|^{\frac{n_*-p-1}{2}} \exp\{-\frac{1}{2} tr \Sigma^{-1}\Sigma_*\} \end{aligned}$$

with c_{n_*} as in (16), being the integration constant of the Wishart distribution. The joint density of Y and (B,Σ^{-1}) in the normal-Wishart model is

$$\begin{aligned} f(Y,B,\Sigma^{-1}) &= (2\pi)^{-\frac{p}{2}(n+K)} |\Sigma|^{-\frac{n+K}{2}} |H_*|^{-\frac{p}{2}} \cdot \\ &\quad \cdot c_{n*}^{-1} |\Sigma_*|^{\frac{n_*}{2}} |\Sigma|^{-\frac{n_*-p-1}{2}} \exp\{-\frac{1}{2} tr \Sigma^{-1}\Sigma_{**}\} \\ &\quad \cdot \exp\{-\frac{1}{2} tr \Sigma^{-1}(B-B_{**})' H_{**}^{-1}(B-B_{**})\}. \end{aligned}$$

Integrating out B and Σ^{-1} gives the result. (q.e.d.)

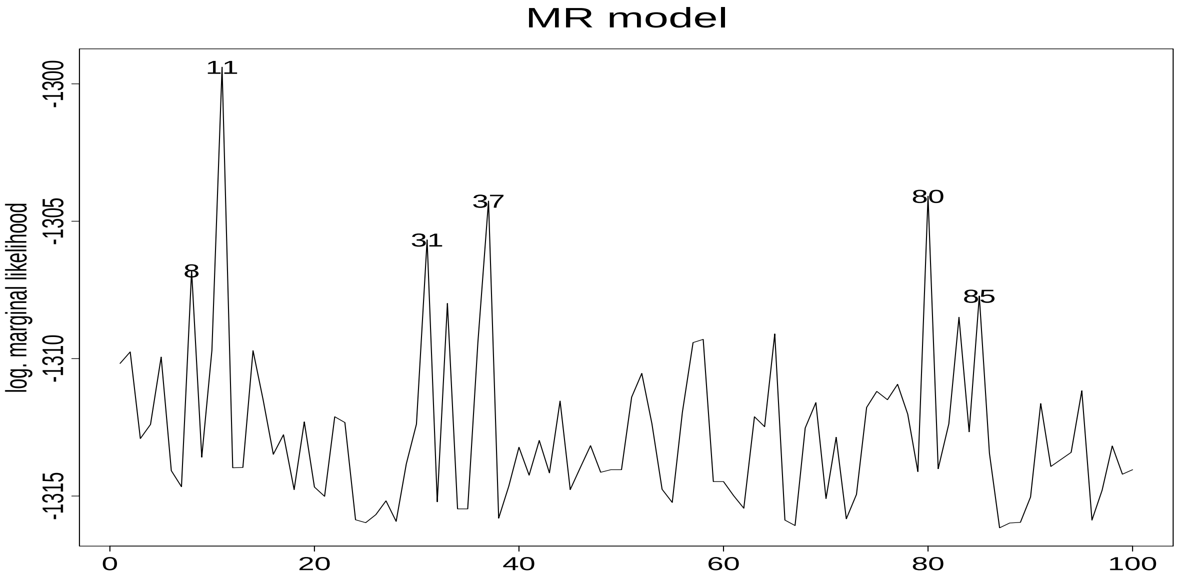
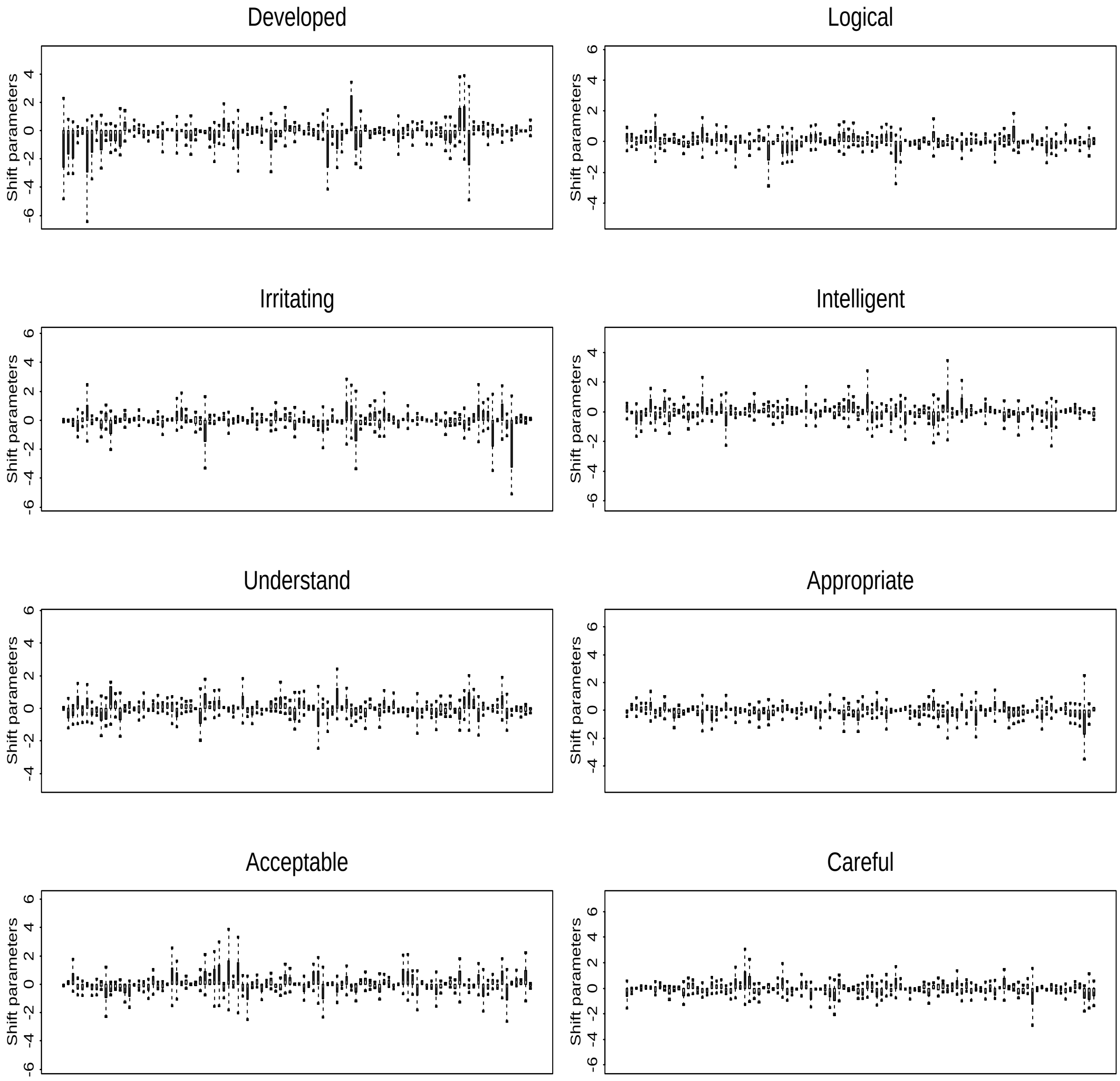


Figure 1: The log. marginal likelihood plot of the language data in Fuller (1987) for the MR outlier model of Polasek and Ren (1998) with fractional prior.



Developed

Logical

Irritating

Intelligent

Understand

Appropriate

Acceptable

Careful

Figure 2: The plot of the shift parameter for potential outliers of the language data in Fuller (1987) in the MR-outlier model.

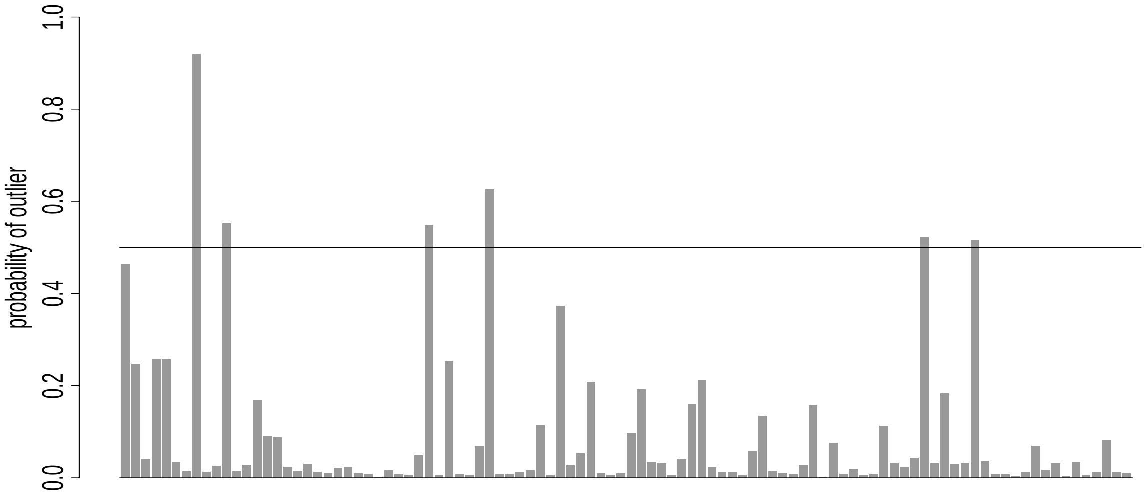


Figure 3: The probability of being an outlier for the language data in Fuller (1987) for the MR-outlier model.

Author: Wolfgang Polasek

Title: Multivariate Regression and ANOVA Models with Outliers: A Comparative Approach

Reihe Ökonomie / Economics Series 136

Editor: Robert M. Kunst (Econometrics)
Associate Editors: Walter Fisher (Macroeconomics), Klaus Ritzberger (Microeconomics)

ISSN: 1605-7996

Stumpergasse 56, A-1060 Vienna • ☎ +43 1 59991-0 • Fax +43 1 59991-555 • http://www.ihs.ac.at

ISSN: 1605-7996