GAUSSIAN SCALE MIXTURE MODELS FOR ROBUST LINEAR MULTIVARIATE REGRESSION WITH MISSING DATA

Juha Ala-Luhtala (Corresponding author)
Department of Mathematics
Tampere University of Technology
PO Box 553, 33101 Tampere, Finland
juha.ala-luhtala@tut.fi

Robert Piché
Department of Automation Science and Engineering
Tampere University of Technology
robert.piche@tut.fi

Short title: ROBUST LINEAR MULTIVARIATE REGRESSION
Key Words: robust linear regression; Gaussian scale mixture; variational Bayes; missing data

ABSTRACT We present an algorithm for multivariate robust Bayesian linear regression with missing data. The iterative algorithm computes an approximative posterior for the model parameters based on the variational Bayes (VB) method. Compared to the EM algorithm, the VB method has the advantage that the variance for the model parameters is also computed directly by the algorithm. We consider three families of Gaussian scale mixture models for the measurements, which include as special cases the multivariate t distribution, the multivariate Laplace distribution, and the contaminated normal model. The observations can contain missing values, assuming that the missing data mechanism can be ignored. A MATLAB/Octave implementation of the algorithm is presented and applied to solve three reference examples from the literature.

1 Introduction

Real data sets often contain extreme observations or outliers, that are not explained by using a normal model for the observations. These outlier observations can have an unduly large influence on the inferences under the normal assumption. There is therefore interest in robust regression, robustness here meaning the tolerance of the model to outliers in the data.

Robust modeling can be based on measurement distributions having fatter tails than the normal distribution. Often used distributions in the statistical literature are Student's t distribution (Blattberg and Gonedes, 1974; Zellner, 1976; West, 1984) and the contaminated normal distribution (Tukey, 1960; Huber, 1964). A robust alternative to ordinary least squares is the method of Least Absolute Deviation (LAD) (Bloomfield and Steiger, 1983), where the absolute value of the errors is minimized instead of the squared errors. The LAD estimate is equivalent to a maximum likelihood estimate using a Laplace distribution for the measurement errors.

A common property of the three fat-tailed distributions mentioned above is that they can be characterised as scale mixtures of normal distributions, also called Gaussian scale mixtures (Andrews and Mallows, 1974). In the Gaussian scale mixture presentation the measurement model is augmented with unobserved weights, so that the conditional distribution of the measurements given the parameters and the weights has a normal distribution. This kind of model enables the use of general algorithms for statistical inference.

Dempster et al. (1977) present the Expectation Maximization (EM) algorithm for Maximum Likelihood (ML) estimation in the so called "incomplete data" models. It is also noted that the EM can be used for computing the posterior mode. Among other examples they consider univariate linear regression with t distributed errors, which they call Iteratively Reweighted Least Squares (IRLS). An extension of IRLS for the multivariate t distribution in linear regression is presented in (Rubin, 2004). Little (1988) studies robust estimation of the multivariate t and contaminated normal models when the observations are allowed to contain missing values. Assuming that the missing data mechanism can be ignored (i.e. the

data are missing at random (MAR)), the EM algorithm can be used to find the ML or MAP estimates of the parameters. Lange et al. (1989) consider multivariate linear and nonlinear regression using the t distribution, where the degrees of freedom is also estimated using the EM algorithm. The EM algorithm for Laplace regression is considered in (Phillips, 2002). Expectation Conditional Maximization (ECM) (Meng and Rubin, 1993) is an extension of the EM algorithm that simplifies the sometimes difficult implementation of the M step. The rate of convergence is improved by the extensions ECME (Liu and Rubin, 1994), Alternating Expectation Conditional Maximization (ACME) (Meng and Dyk, 1997) and Parameter Expanded Expectation Maximization (PX-EM) (Liu et al., 1998). The EM algorithm does not give directly information about the reliability of the parameter estimates. This can be addressed however by using e.g. asymptotic results (Meng and Rubin, 1991) or bootstrapping (McLachlan and Krishnan, 2008, pp. 130-131).

In Bayesian statistical inference we are interested in the full posterior of the model parameters. Nowadays Bayesian analysis is done mostly by Markov Chain Monte Carlo (MCMC) methods, such as the Gibbs sampler, which iteratively produce samples from the full posterior. Verdinelli and Wasserman (1991) consider Bayesian analysis of the univariate Student t and contaminated (location-shift) normal models using the Gibbs sampler. The implementation for the Student t distribution makes use of the Gaussian scale mixture presentation. Liu (1996) uses a Monte Carlo method called Data Augmentation (DA) (Tanner and Wong, 1987) for multivariate robust linear regression with missing data using the multivariate t, the contaminated normal, and the slash distribution. The algorithm makes use of the Gaussian scale mixture presentation for all the distributions. The DA and Gibbs sampler algorithms can be viewed as stochastic extensions of the EM and ECM algorithms respectively.

An alternative for the computationally heavy Monte Carlo methods is provided by an approximate method called variational Bayes (VB). In the variational Bayesian EM (VB-EM), the intractable posterior is approximated by assuming that it factorizes between model parameters and latent variables (Beal and Ghahramani, 2003; Beal, 2003). The VB-EM algorithm iteratively minimizes the Kullback-Leibler divergence between the true posterior

and the approximate distribution. The VB-EM algorithm reduces to the EM algorithm when the approximate distribution for the parameters is assumed to be a Dirac delta function (Beal and Ghahramani, 2003). Titterington (2004) provides more discussion about the VB-EM algorithm, especially in the neural networks point of view. Tipping and Lawrence (2003) use the variational approximation for robust linear interpolation using the t distribution. Penny et al. (2007) study the univariate linear regression model with observations from the contaminated normal distribution. Wand et al. (2011) considers variational inference using the mean field method for several statistical distributions, including the Student t model for univariate robust regression. More examples on variational inference is provided in (Ormerod and Wand, 2010), where e.g. Baeysian logistic regression using variational inference is presented. Also, it is shown how the mean field variational approximation is connected with the MCMC method Gibbs sampling. The VB approach has also been studied in connection with robust autoregressive modeling (Roberts and Penny, 2002) and nonlinear regression (Chappell et al., 2009). See also VIBES (Bishop et al., 2002), a VB-based software package for statistical inference with Bayesian networks.

The need for robust methods for statistical analysis is recognized in many statistical software packages. LIBRA (Verboven and Hubert, 2005) is a library of MATLAB functions implementing many robust statistical methods, although not Bayesian regression. The `monomvn` package for R provides a Bayesian treatment of robust linear regression with missing data using the t-distribution; the computations are based on Gibbs sampler and DA algorithms.

Much of the literature on robust inference is concentrated on using the t distribution as a robust alternative for the normal distribution. Some authors also point out the use of Laplace or finite mixtures of normal distributions as other robust choices. Many of the referenced inference methods make use of the Gaussian scale mixture presentation of Andrews and Mallows (1974), especially for the t distribution. In this paper we present an algorithm for robust multivariate linear regression with missing data using a general Gaussian scale mixture family of distributions. The multivariate t, multivariate Laplace and multivariate

contaminated normal distributions are included as special cases. In the case of missing data we assume that the missing data mechanism can be ignored as described in (Little and Rubin, 2002). The VB method provides an unified treatment of all the different models and is used to compute an approximation for the posterior distribution of the model parameters.

The rest of the paper is organised as follows. Section 2 presents the statistical model and the different distributions used for robust modeling. In section 3 the VB method is presented and the equations needed for computing the approximate posterior are derived. The algorithm and its implementation in MATLAB/Octave are also presented here. The usage of the MATLAB/Octave implementation is presented in section 4 with three examples from the literature.

2 Robust linear regression model

The sampling model considered in this paper is

$$\mathbf{y}_n \,|\, \mathbf{x}, \mathbf{Q}, w_n \sim \text{Normal}(\mathbf{H}_n\mathbf{x}, \mathbf{Q}/w_n), \tag{1}$$

where $\mathbf{y}_n$ is a d-dimensional observation vector, $\mathbf{H}_n$ is a $d \times p$ design matrix, $\mathbf{x}$ is a p-dimensional vector of parameters, w_n is a positive scalar, and $\mathbf{Q}$ is a $d \times d$ symmetric positive definite matrix. The N observations $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are assumed to be conditionally independent given the model parameters $\mathbf{x}, \mathbf{Q}, w_n$. The scale parameter w_n is assumed to be independent of $\mathbf{x}, \mathbf{Q}$.

Marginalisation of w_n gives the sampling model in the form

$$\begin{aligned} p(\mathbf{y}_n \,|\, \mathbf{x}, \mathbf{Q}) &= \int p(\mathbf{y}_n, w_n \,|\, \mathbf{x}, \mathbf{Q}) \, \mathrm{d}w_n = \int p(\mathbf{y}_n \,|\, \mathbf{x}, \mathbf{Q}, w_n) p(w_n) \, \mathrm{d}w_n \\ &= \int \text{Normal}(\mathbf{y}_n; \mathbf{H}_n\mathbf{x}, \mathbf{Q}/w_n) p(w_n) \, \mathrm{d}w_n. \end{aligned} \tag{2}$$

Sampling models obtained using Eq. (2) are called Gaussian scale mixtures (Andrews and Mallows, 1974). Robust regression can be achieved by choosing the prior distribution $p(w_n)$ such that (2) has fatter tails than the corresponding normal distribution. The mean and variance for the distribution in Eq. (2) can be found using formulas for conditional expectation

and variance (Gelman et al., 2003, p.37). The mean is given by

$$\mathbb{E}(\mathbf{y}_n \,|\, \mathbf{x}, \mathbf{Q}) = \mathbb{E}(\mathbb{E}(\mathbf{y} \,|\, \mathbf{x}, \mathbf{Q}, w_n)) = \mathbf{H}_n\mathbf{x} \tag{3}$$

and the variance is found by

$$\begin{aligned}\mathrm{Var}(\mathbf{y}_n \,|\, \mathbf{x}, \mathbf{Q}) &= \mathbb{E}(\mathrm{Var}(\mathbf{y} \,|\, \mathbf{x}, \mathbf{Q}, w_n)) + \mathrm{Var}(\mathbb{E}(\mathbf{y} \,|\, \mathbf{x}, \mathbf{Q}, w_n)) \\ &= \mathbb{E}(w_n^{-1})\mathbf{Q} \end{aligned} \tag{4}$$

We consider three families of prior distributions for w_n.

1. A gamma distribution, where the density takes the form

$$p(w_n) \propto w_n^{\alpha-1} e^{-\beta w_n}, \quad w_n > 0. \tag{5}$$

   The variance for the observation is given by

$$\mathrm{Var}(\mathbf{y}_n \,|\, \mathbf{x}, \mathbf{Q}) = \frac{\beta}{\alpha - 1}\mathbf{Q}. \tag{6}$$

   In the case $\alpha = \beta = \nu/2$, the random variable w_n has a chi-squared distribution, and the observations have a multivariate t distribution. The observations' distribution in the general case is known as the generalised t distribution of Arellano-Valle and Bolfarine (Kotz and Nadarajah, 2004, p.94).

2. An inverse gamma distribution, with density

$$p(w_n) \propto w_n^{-\alpha-1} e^{-\beta w_n^{-1}}, \quad w_n > 0. \tag{7}$$

   The variance for the observation is given by

$$\mathrm{Var}(\mathbf{y}_n \,|\, \mathbf{x}, \mathbf{Q}) = \frac{\alpha}{\beta}\mathbf{Q}. \tag{8}$$

   In the case $\alpha = \beta = 1$, the random variable w_n^{-1} has a standard exponential distribution, and the observations have a multivariate symmetric Laplace distribution (Kotz et al., 2001, p.246).

3. A two-component Gaussian mixture, with density

$$p(w_n) = (1-\epsilon)\delta(w_n - 1) + \epsilon\delta(w_n - 1/c). \tag{9}$$

This gives the "contaminated normal" observation model introduced by Tukey (Tukey, 1960)

$$p(\mathbf{y}_n \,|\, \mathbf{x}, \mathbf{Q}) = (1-\epsilon)\,\mathrm{Normal}(\mathbf{y}_n; \mathbf{H}_n\mathbf{x}, \mathbf{Q}) + \epsilon\,\mathrm{Normal}(\mathbf{y}_n; \mathbf{H}_n\mathbf{x}, c\mathbf{Q}), \tag{10}$$

where $0 < \epsilon < 1$ is the probability of getting an outlier and the factor $c > 1$ is used to model the larger variance of the outliers.

Method to estimate the hyperparameters, for the case that these are unknown, is given in Section 3.2. However, this generally requires a lot of data to be useful. Also, the estimation of the factor c in the contaminated normal model can not be included in the variational Bayes algorithm considered in this paper. However, as described in Section 3.2, a grid search based method can be used instead.

The observations are allowed to contain missing data elements. We consider the case when data is missing at random (MAR) as described in (Little and Rubin, 2002, p. 12). Under the MAR assumption we can ignore the missing data mechanism and base our inference solely on the observed data. For each observation $\mathbf{y}_n$ define a permutation matrix $\mathbf{M}_n = \begin{bmatrix} \mathbf{M}_n^{\mathrm{obs}} & \mathbf{M}_n^{\mathrm{miss}} \end{bmatrix}$ such that

$$\mathbf{y}_n = \mathbf{M}_n \begin{bmatrix} \mathbf{y}_n^{\mathrm{obs}} \\ \mathbf{y}_n^{\mathrm{miss}} \end{bmatrix}, \tag{11}$$

where the d_n vector $\mathbf{y}_n^{\mathrm{obs}}$ and $d-d_n$ vector $\mathbf{y}_n^{\mathrm{miss}}$ are the observed and missing part respectively. If $\mathbf{y}_n$ has no missing values, take $\mathbf{M}_n = \mathbf{I}$. Data vectors with all the elements missing are discarded, since under the assumption of randomly missing data, these do not contain any useful information. The sets $\mathbf{Y}^{\mathrm{obs}}$ and $\mathbf{Y}^{\mathrm{miss}}$ denote all the observed and missing data respectively.

We take independent priors for the parameters $p(\mathbf{x}, \mathbf{Q}) = p(\mathbf{x})p(\mathbf{Q})$, where

$$p(\mathbf{x}) \propto 1 \tag{12}$$

and

$$p(\mathbf{Q}) \propto |\mathbf{Q}|^{-(m+d+1)/2} \exp\left[-\tfrac{1}{2}\operatorname{tr}(\mathbf{Q}^{-1}\mathbf{A})\right]. \tag{13}$$

The distribution $p(\mathbf{Q})$ is the Inverse Wishart distribution, where m and $\mathbf{A}$ are parameters. Taking $m = 0$ and $\mathbf{A} = 0$ we get the noninformative Jeffreys prior (Gelman et al., 2003).

3 Approximate Bayesian inference

3.1 Theory

The solution of the linear regression problem is the posterior distribution $p(\mathbf{x}, \mathbf{Q} \mid \mathbf{Y}^{\text{obs}})$. This distribution can be obtained by integrating out the latent variables and the missing data from the full posterior

$$p(\mathbf{x}, \mathbf{Q} \mid \mathbf{Y}^{\text{obs}}) = \iint p(\mathbf{x}, \mathbf{Q}, \mathbf{w}, \mathbf{Y}^{\text{miss}} \mid \mathbf{Y}^{\text{obs}}) \, \mathrm{d}\mathbf{w} \, \mathrm{d}\mathbf{Y}^{\text{miss}}. \tag{14}$$

In this paper we consider a Variational Bayes (VB) approximation to the posterior distribution. The variational approach in general is based on maximizing the variational lower bound of the logarithm of the marginal likelihood (Bishop, 2006; MacKay, 2003; Beal and Ghahramani, 2003; Beal, 2003)

$$\log p(\mathbf{Y}^{\text{obs}}) \geq \int q(\mathbf{x}, \mathbf{Q}, \mathbf{w}, \mathbf{Y}^{\text{miss}}) \log \frac{p(\mathbf{Y}^{\text{obs}}, \mathbf{x}, \mathbf{Q}, \mathbf{w}, \mathbf{Y}^{\text{miss}})}{q(\mathbf{x}, \mathbf{Q}, \mathbf{w}, \mathbf{Y}^{\text{miss}})} \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{Q} \, \mathrm{d}\mathbf{w} \, \mathrm{d}\mathbf{Y}^{\text{miss}}, \tag{15}$$

where $q(\mathbf{x}, \mathbf{Q}, \mathbf{w}, \mathbf{Y}^{\text{miss}})$ is any distribution over the latent variables and model parameters. It can be shown that maximizing the lower bound is equivalent to minimizing the Kullback-Leibler divergence between $q(\mathbf{x}, \mathbf{Q}, \mathbf{w}, \mathbf{Y}^{\text{miss}})$ and the full posterior:

$$\mathrm{KL}(q \,\|\, p) = \int q(\mathbf{x}, \mathbf{Q}, \mathbf{w}, \mathbf{Y}^{\text{miss}}) \log \frac{q(\mathbf{x}, \mathbf{Q}, \mathbf{w}, \mathbf{Y}^{\text{miss}})}{p(\mathbf{x}, \mathbf{Q}, \mathbf{w}, \mathbf{Y}^{\text{miss}} \mid \mathbf{Y}^{\text{obs}})} \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{Q} \, \mathrm{d}\mathbf{w} \, \mathrm{d}\mathbf{Y}^{\text{miss}}. \tag{16}$$

In the VB approach we make a fully factorized approximation for the posterior distribution (Bishop, 2006; MacKay, 2003)

$$\begin{aligned} p(\mathbf{x}, \mathbf{Q}, \mathbf{w}, \mathbf{Y}^{\text{miss}} \mid \mathbf{Y}^{\text{obs}}) &\approx q(\mathbf{x}) q(\mathbf{Q}) q(\mathbf{w}) q(\mathbf{Y}^{\text{miss}}) \\ &= q(\mathbf{x}) q(\mathbf{Q}) \prod_{n=1}^{N} q(w_n) q(\mathbf{y}_n^{\text{miss}}). \end{aligned} \tag{17}$$

That is, the posterior model parameters and latent variables are approximated as being mutually independent. Note that the functional form of the approximating distributions $q(\cdot)$ is not fixed. Generally the approximations from variational inference tend to be more compact than the true distribution (MacKay, 2003, p. 431). It can be shown (Bishop, 2006) that the optimal distributions in the sense of KL-divergence satisfy the equations

$$\log q(s) = \mathbb{E}_{-s} \log p(\mathbf{x}, \mathbf{Q}, \mathbf{w}, \mathbf{Y}^{\text{miss}}, \mathbf{Y}^{\text{obs}}), \tag{18}$$

where

$$s \in \{\mathbf{x}, \mathbf{Q}, w_1, \ldots, w_N, \mathbf{y}_1^{\text{miss}}, \ldots, \mathbf{y}_N^{\text{miss}}\}. \tag{19}$$

The notation $\mathbb{E}_{-s}(\cdot)$ means that the expectation is taken with respect to all variables other than s. The parameters of the approximation can be found by fixed-point iteration based on Eq. (18).

We next derive the equations for computing the approximating distributions using Eq. (18). The log-probability of the joint distribution of all the variables is

$$\begin{aligned}
&\log p(\mathbf{x}, \mathbf{Q}, \mathbf{w}, \mathbf{Y}^{\text{miss}}, \mathbf{Y}^{\text{obs}}) = \\
&\quad -\frac{N+m+d+1}{2} \log |\mathbf{Q}| - \frac{1}{2} \operatorname{tr}(\mathbf{Q}^{-1}\mathbf{A}) \\
&\quad \sum_{n=1}^{N} \left[-\frac{w_n}{2} (\mathbf{y}_n - \mathbf{H}_n \mathbf{x})^T \mathbf{Q}^{-1} (\mathbf{y}_n - \mathbf{H}_n \mathbf{x}) + \frac{d}{2} \log w_n + \log p(w_n) \right] + \text{const},
\end{aligned} \tag{20}$$

where $\mathbf{y}_n$ is of the form in Eq. (11). Taking expectation with respect to all other variables than $\mathbf{x}$ and absorbing all the terms that do not involve $\mathbf{x}$ into the constant term, we get

$$\log q(\mathbf{x}) = \sum_{n=1}^{N} -\frac{\bar{w}_n}{2} \left[\mathbf{x}^T \mathbf{H}_n^T \mathbf{S} \mathbf{H}_n \mathbf{x} - 2\mathbf{x}^T \mathbf{H}_n^T \mathbf{S} \bar{\mathbf{y}}_n) \right] + \text{const}, \tag{21}$$

where $\bar{w}_n = \mathbb{E}(w_n)$, $\mathbf{S} = \mathbb{E}(\mathbf{Q}^{-1})$ and

$$\bar{\mathbf{y}}_n = \mathbf{M}_n \begin{bmatrix} \mathbf{y}_n^{\text{obs}} \\ \mathbb{E}(\mathbf{y}_n^{\text{miss}}) \end{bmatrix}. \tag{22}$$

The expression in Eq. (21) is recognized as the logarithm of the normal density with variance

$$\mathbf{P} = \left(\sum_{n=1}^{N} \bar{w}_n \mathbf{H}_n^T \mathbf{S} \mathbf{H}_n \right)^{-1} \tag{23}$$

and mean

$$\bar{\mathbf{x}} = \mathbf{P}\left(\sum_{n=1}^{N} \bar{w}_n \mathbf{H}_n^T \mathbf{S} \bar{\mathbf{y}}_n\right). \tag{24}$$

To find the distribution for $\mathbf{Q}$, we first note that the quadratic form in Eq. (48) can be written using the matrix trace operator as

$$(\mathbf{y}_n - \mathbf{H}_n\mathbf{x})^T \mathbf{Q}^{-1} (\mathbf{y}_n - \mathbf{H}_n\mathbf{x}) = \mathrm{tr}\left[\mathbf{Q}^{-1}(\mathbf{y}_n - \mathbf{H}_n\mathbf{x})(\mathbf{y}_n - \mathbf{H}_n\mathbf{x})^T\right]. \tag{25}$$

Using this result and taking again expectation with respect to all other variables than $\mathbf{Q}$, we get

$$\log q(\mathbf{Q}) = -\frac{1}{2}\,\mathrm{tr}\left[\mathbf{Q}^{-1}\left(\mathbf{A} + \mathbf{R}\right)\right] - \frac{N + m + d + 1}{2}\log|\mathbf{Q}| + \text{const}, \tag{26}$$

where

$$\mathbf{R} = \sum_{n=1}^{N} \bar{w}_n \left[(\bar{\mathbf{y}}_n - \mathbf{H}_n\bar{\mathbf{x}})(\bar{\mathbf{y}}_n - \mathbf{H}_n\bar{\mathbf{x}})^T + \Sigma_n + \mathbf{H}_n\mathbf{P}\mathbf{H}_n^T\right] \tag{27}$$

and

$$\Sigma_n = \mathbf{M}_n \begin{bmatrix} \underline{0} & \underline{0} \\ \underline{0} & \mathrm{Cov}(\mathbf{y}_n^{\mathrm{miss}}) \end{bmatrix} \mathbf{M}_n^T. \tag{28}$$

This is recognized as the inverse Wishart distribution with degrees of freedom $N + m$ and scale matrix $\mathbf{A} + \mathbf{R}$. From this it follows that $\mathbf{Q}^{-1}$ has Wishart distribution with the same degrees of freedom and scale matrix $(\mathbf{A} + \mathbf{R})^{-1}$. Using this result we can compute

$$\mathbf{S} = \mathbb{E}(\mathbf{Q}^{-1}) = (N + m)(\mathbf{A} + \mathbf{R})^{-1}. \tag{29}$$

For one dimensional measurements (i.e. $d = 1$) the inverse Wishart distribution reduces to the inverse Gamma distribution.

For the missing values we find that

$$\log q(\mathbf{y}_n^{\mathrm{miss}}) = -\frac{\bar{w}_n}{2}\left[\mathbf{y}_n^T\mathbf{S}\mathbf{y}_n - 2\mathbf{y}_n^T\mathbf{S}\mathbf{H}_n\bar{\mathbf{x}}\right] + \text{const}. \tag{30}$$

Inserting Eq. (11) for $\mathbf{y}_n$, we find that the distributions for the missing values are normal with means

$$\mathbb{E}(\mathbf{y}_n^{\mathrm{miss}}) = (\mathbf{M}_n^{\mathrm{miss}})^T\mathbf{H}_n\bar{\mathbf{x}} + (\Sigma_n^{\mathrm{om}})^T\left(\Sigma_n^{\mathrm{obs}}\right)^{-1}(\mathbf{y}_n^{\mathrm{obs}} - \mathbf{M}_n^{\mathrm{obs}}\mathbf{H}_n\bar{\mathbf{x}}) \tag{31}$$

and variance

$$\mathrm{Cov}(\mathbf{y}_n^{\mathrm{miss}}) = \Sigma_n^{\mathrm{miss}} - (\Sigma_n^{\mathrm{om}})^T \left(\Sigma_n^{\mathrm{obs}}\right)^{-1} \Sigma_n^{\mathrm{om}}, \tag{32}$$

where

$$\Sigma_n^{\mathrm{obs}} = (\mathbf{M}_n^{\mathrm{obs}})^T [\bar{w}_n \mathbf{S}]^{-1} \mathbf{M}_n^{\mathrm{obs}} \tag{33}$$

$$\Sigma_n^{\mathrm{miss}} = (\mathbf{M}_n^{\mathrm{miss}})^T [\bar{w}_n \mathbf{S}]^{-1} \mathbf{M}_n^{\mathrm{miss}} \tag{34}$$

$$\Sigma_n^{\mathrm{om}} = (\mathbf{M}_n^{\mathrm{obs}})^T [\bar{w}_n \mathbf{S}]^{-1} \mathbf{M}_n^{\mathrm{miss}} \tag{35}$$

The form of the optimal distribution for weights w_n depends on the prior distribution $p(w_n)$

$$q(w_n) \propto w_n^{d/2} \exp\left(-\frac{w_n}{2} l_n\right) p(w_n), \tag{36}$$

where

$$l_n = (\bar{\mathbf{y}}_n - \mathbf{H}_n \mathbf{m})^T \mathbf{S} (\bar{\mathbf{y}}_n - \mathbf{H}_n \mathbf{m}) + \mathrm{tr}[\mathbf{S}(\mathbf{H}_n \mathbf{P} \mathbf{H}_n^T + \Sigma_n)]. \tag{37}$$

The distributions for the three families of priors considered in this work are as follows.

1. For a-priori gamma distributed w_n the optimal distribution is also gamma distribution with density

   $$q(w_n) \propto w_n^{\alpha + d/2 - 1} e^{-w_n(\beta + l_n/2)} \tag{38}$$

   and mean

   $$\bar{w}_n = \frac{\alpha + d/2}{\beta + l_n/2}. \tag{39}$$

2. For a priori inverse gamma distributed w_n the optimal distribution is

   $$q(w_n) \propto w_n^{d/2 - \alpha - 1} e^{-\frac{1}{2}(l_n w_n + 2\beta w_n^{-1})}. \tag{40}$$

   This is recognized as a Generalized Inverse Gaussian (GIG) distribution (Jørgensen, 1982). The GIG distribution has mean

   $$\bar{w}_n = \sqrt{\frac{2\beta}{l_n}} \frac{K_{d/2-\alpha+1}(\sqrt{2\beta l_n})}{K_{d/2-\alpha}(\sqrt{2\beta l_n})}, \tag{41}$$

   where K_p is the modified Bessel function of the second kind with order p.

3. For the discrete distribution of Eq. (9) the optimal distribution is the discrete distribution

$$q(w_n) \propto (1-\epsilon)e^{-\frac{l_n}{2}}\delta(w_n - 1) + \epsilon c^{-d/2}e^{-\frac{l_n}{2c}}\delta(w_n - 1/c), \tag{42}$$

with mean

$$\bar{w}_n = \frac{(1-\epsilon)e^{-\frac{l_n}{2}} + \epsilon c^{-d/2-1}e^{-\frac{l_n}{2c}}}{(1-\epsilon)e^{-\frac{l_n}{2}} + \epsilon c^{-d/2}e^{-\frac{l_n}{2c}}}. \tag{43}$$

3.1.1 Computation of the variational lower bound

The lower bound of the marginal likelihood defined in Eq. (15) should be non-decreasing during the iterative variational inference algorithm (Bishop, 2006, p. 481) and can be used to check the convergence of the algorithm. Expanding the expression for the lower bound we have

$$\begin{aligned} L(q) = {} & \mathbb{E}\log p(\mathbf{x}, \mathbf{Q}, \mathbf{w}, \mathbf{Y}^{\text{miss}}, \mathbf{Y}^{\text{obs}}) - \mathbb{E}\log q(\mathbf{x}) - \mathbb{E}\log q(\mathbf{Q}) + \sum_{n=1}^{N} - \mathbb{E}\log q(w_n) \\ & + \sum_{n=1}^{N} - \mathbb{E}\log q(\mathbf{y}_n^{\text{miss}}), \end{aligned} \tag{44}$$

where the first expectation is with respect to the whole approximate posterior in Eq. (17) and the rest with respect to the corresponding approximate marginal distributions. The terms of the form $-\mathbb{E}\log q(s)$ are recognized as the differential entropies of the corresponding distributions. The entropies for $q(\mathbf{x})$, $q(\mathbf{Q})$ and $q(\mathbf{y}_n^{\text{miss}})$ are computed using equations (Bishop, 2006, pp. 685–693)

$$-\mathbb{E}\log q(\mathbf{x}) = \frac{1}{2}\log|\mathbf{P}| + \frac{p}{2}\left[1 + \log(2\pi)\right] \tag{45}$$

$$\begin{aligned} -\mathbb{E}\log q(\mathbf{Q}) = {} & \frac{d+1}{2}\log|\mathbf{A} + \mathbf{R}| + \frac{d(N+m)}{2} + \\ & \log\left[2^{d(N+m)/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma\left(\frac{N+m+(1-j)}{2}\right)\right] \end{aligned} \tag{46}$$

$$-\mathbb{E}\log q(\mathbf{y}^{\text{miss}}) = \frac{1}{2}\log|\operatorname{Cov}(\mathbf{y}_n^{\text{miss}})| + \frac{d_n}{2}\left[1 + \log(2\pi)\right]. \tag{47}$$

The expression for $\mathbf{P}$, $\mathbf{R}$ and $\mathrm{Cov}(\mathbf{y}_n^{\mathrm{miss}})$ are given by Eqs. (23), (27) and (32) respectively. The entropy of the inverse Wishart distribution can be derived using entropy of the Wishart distribution.

Inserting the expression in Eq. (48) for the first term in the lower bound and evaluating the expectations, we get

$$\begin{aligned}\mathbb{E}\log p(\mathbf{x},\mathbf{Q},\mathbf{w},\mathbf{Y}^{\mathrm{miss}},\mathbf{Y}^{\mathrm{obs}}) = \\ -\frac{N+m+d+1}{2}\log|\mathbf{A}+\mathbf{R}| - \frac{1}{2}\operatorname{tr}(\mathbf{S}\mathbf{A}) \\ \sum_{n=1}^{N}\left[-\frac{\bar{w}_n}{2}l_n + \frac{d}{2}\mathbb{E}\log w_n + \mathbb{E}\log p(w_n)\right] + \mathrm{const},\end{aligned} \tag{48}$$

where the terms $\mathbf{R}$, $\mathbf{S}$ and l_n are given by Eqs. (27), (29) and (37) respectively. The constant term in the expression, i.e. the term that does not change during iterations, is given by

$$\mathrm{const} = \frac{m}{2}\log|\mathbf{A}| - \log\left[2^{md/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma\left(\frac{m+1-j}{2}\right)\right] - \frac{d}{2}\log(2\pi). \tag{49}$$

The expressions for the terms depending on the prior distribution of w_n are given as follows:

1. For a-priori gamma distributed w_n the optimal distribution is the gamma distribution $w_n \sim \mathrm{Gamma}(a_n, b_n)$, with parameters $a_n = \alpha + d/2$ and $b_n = \beta + l_n/2$. The mean $\bar{w}_n$ is given by Eq. (39). The log-expectation and the entropy are given by (Bishop, 2006, p. 688)

$$\mathbb{E}\log w_n = \psi(a_n) - \log b_n \tag{50}$$

and

$$-\mathbb{E}\log q(w_n) = \log\Gamma(a_n) - (a_n-1)\psi(a_n) - \log b_n + a_n, . \tag{51}$$

where $\psi(x)$ is the digamma function. The expectation of the prior log-probability is given by

$$\mathbb{E}\log p(w_n) = \alpha\log\beta - \log\Gamma(\alpha) + (\alpha-1)\,\mathbb{E}\log w_n - \beta\bar{w}_n. \tag{52}$$

2. For a-priori inverse gamma distributed w_n the optimal distribution is the GIG distribution $w_n \sim \mathrm{GIG}(p_n, a_n, b_n)$, with parameters $p_n = d/2 - \alpha$, $a_n = l_n$ and $b_n = 2\beta$. The mean $\bar{w}_n$ is given by Eq. (41). The log-expectation is given by (Jørgensen, 1982, p. 21)

$$\mathbb{E} \log w_n = \log \left(\sqrt{\frac{b_n}{a_n}} \right) + \frac{\left[\frac{\partial}{\partial v} K_v(\sqrt{a_n b_n}) \right]_{v=p_n}}{K_p(\sqrt{a_n b_n})}. \tag{53}$$

The partial derivative of the modified Bessel function with respect to the order v can be evaluated using formulas from (Abramowitz and Stegun, 1965, p. 377). The entropy is given by

$$-\mathbb{E} \log q(w_n) = \log \left[\frac{(a_n/b_n)^{p_n/2}}{2K_{p_n}(\sqrt{a_n b_n})} \right] + (p-1)\,\mathbb{E} \log w_n - \frac{1}{2} \left(a_n \bar{w}_n + b_n\, \mathbb{E}\, w_n^{-1} \right), \tag{54}$$

where

$$\mathbb{E}\, w_n^{-1} = \sqrt{\frac{a_n}{b_n}} \frac{K_{p-1}(\sqrt{a_n b_n})}{K_{p_n}(\sqrt{a_n b_n})}. \tag{55}$$

The expectation of the prior log-probability is given by

$$\mathbb{E} \log p(w_n) = \alpha \log \beta - \log \Gamma(\alpha) - (\alpha + 1)\,\mathbb{E} \log w_n - \beta\, \mathbb{E}\, w_n^{-1}. \tag{56}$$

3. For the discrete distribution of Eq. (42) the mean $\bar{w}_n$ is given by Eq. (43). The log-expectation and entropy are given by

$$\mathbb{E} \log w_n = \log \left(\frac{1}{c} \right) q(w_n = 1/c) \tag{57}$$

and

$$-\mathbb{E} \log q(w_n) = -q(w_n = 1) \log q(w_n = 1) - q(w_n = 1/c) \log q(w_n = 1/c), \tag{58}$$

where $q(w_n = s)$ is the value of the pdf $q(w_n)$ evaluated at point s. The expectation of the prior log-probability is given by

$$\mathbb{E} \log p(w_n) = q(w_n = 1) \log(1 - \epsilon) + q(w_n = 1/c) \log(\epsilon). \tag{59}$$

Collecting the results, the expression for the lower bound is given by

$$L(q) = -\frac{N+m}{2}\log|\mathbf{A}+\mathbf{R}| + \frac{1}{2}\log|\mathbf{P}| + \frac{1}{2}\sum_{n=1}^{N}\log|\operatorname{Cov}(\mathbf{y}_n^{\text{miss}})| - \frac{1}{2}\operatorname{tr}(\mathbf{SA})$$
$$\sum_{n=1}^{N}\left[-\frac{\bar{w}_n}{2}l_n + \frac{d}{2}\mathbb{E}\log w_n + \mathbb{E}\log p(w_n)\right] + \text{const}, \tag{60}$$

where $\bar{w}_n$, $\mathbb{E}\log w_n$ and $\mathbb{E}\log p(w_n)$ are computed using results depending on the used prior density $p(w_n)$, and the constant term is given by

$$\text{const} = \frac{p + d(N+m) + \sum_{n=1}^{N} d_n)}{2} + \log(2\pi)\frac{p - d + \sum_{n=1}^{N} d_n}{2}$$
$$+ \log\left[2^{dN/2}\prod_{j=1}^{d}\frac{\Gamma\left(\frac{N+m+(1-j)}{2}\right)}{\Gamma\left(\frac{m+1-j}{2}\right)}\right] + \frac{m}{2}\log|\mathbf{A}|. \tag{61}$$

3.1.2 Estimating the hyperparameters

The hyperparameters of the prior distribution for the weights $p(w_n)$ might be unknown. To estimate also the hyperparameters, an additional maximization step can be included to the variational Bayes algorithm as described by Beal (Beal, 2003, pp. 61-62). The iteration for the hyperparameter estimation proceeds as follows. Given the previous estimates for the hyperparameters, the optimal distributions for the variables $\mathbf{x}$, $\mathbf{Q}$, $\mathbf{w}$ and $\mathbf{Y}^{\text{miss}}$ are computed. After this, the obtained optimal distributions are kept unchanged and the lower bound is maximized with respect to the hyperparameters. For the Gaussian scale mixture models used in this paper, the maximization of the lower bound with respect to the hyperparameters reduces to maximizing the sum of the expected log-probabilites of the prior distribution:

$$\sum_{n=1}^{N}\mathbb{E}\log p(w_n). \tag{62}$$

For the three different families of prior distributions $p(w_n)$ the maximization proceeds as follows.

1. For a-priori gamma distributed w_n, the Eq. (52) is maximized with respect to the hyperparameters α and β. The terms $\mathbb{E}\log w_n$ and $\bar{w}_n$ are given by Eqs. (50) and (39),

where the previous estimated values are used for the hyperparameters α and β. Taking derivatives with respect to α and β, and setting them to zero, we get the equations

$$N \log \beta - N\Psi(\alpha) + \sum_{n=1}^{N} \mathbb{E} \log(w_n) = 0 \tag{63}$$

$$N\frac{\alpha}{\beta} - \sum_{n=1}^{N} \bar{w}_n = 0, \tag{64}$$

where

$$\Psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \tag{65}$$

is the digamma function. There is no explicit solution, but an iterative method (e.g. the Newton method) can be used to solve the equations. The Hessian is given by

$$\mathbf{H} = \begin{bmatrix} -N\Psi'(\alpha) & \frac{N}{\beta} \\ \frac{N}{\beta} & -N\frac{\alpha}{\beta}^2 \end{bmatrix}. \tag{66}$$

The derivative of the digamma function $\Psi'(\alpha)$ is also known as the trigamma function.

2. For a-priori inverse gamma distributed w_n, the Eq. (56) is maximized with respect to the hyperparameters α and β. The terms $\mathbb{E} \log w_n$ and $\mathbb{E}\, w_n^{-1}$ are computed using equations (53) and (55), where values of the previous estimates are used for α and β. Taking derivatives with respect to α and β and evaluating to zero, gives equations

$$N \log \beta - N\Psi(\alpha) - \sum_{n=1}^{N} \mathbb{E} \log(w_n) = 0 \tag{67}$$

$$N\frac{\alpha}{\beta} - \sum_{n=1}^{N} \mathbb{E}\, w_n^{-1} = 0. \tag{68}$$

Iterative method can be used to solve the obtained equations. The Hessian is the same as for the a-priori gamma distributed w_n.

3. For the discrete distribution of Eq. (42) the hyperparameters are the probability of the outlier ϵ and the scaling factor c. Derivating Eq. (59) with respect to ϵ and setting the derivative to zero, gives

$$\epsilon = \frac{q(w_n = 1)}{q(w_n = 1) + q(w_n = 1/c)}. \tag{69}$$

Note that this method can not be used to estimate the scaling factor, since c appears only as the argument of the discrete distribution q. The scaling factor c can be estimated using a grid search based method, where the VBEM algorithm is run for several different values of c and the best estimate is chosen to be the value of c for which the final value of the lower bound is largest.

3.2 Algorithm

The VB algorithm for the robust regression proceeds iteratively starting from some initial guess for the statistics $\bar{\mathbf{x}}$, $\mathbf{P}$, $\mathbf{S}$ and $\bar{w}_1, \ldots, \bar{w}_N$. We first update $\bar{\mathbf{y}}_n$ and Σ_n for $n = 1, \ldots, N$ using Eqs. (31), (22), (32) and (28). Next we use these values and update $\bar{\mathbf{x}}$, $\mathbf{P}$ and $\mathbf{S}$ using Eqs. (24), (23), (27) and (29). Last we update the means for the weights w_n using the appropriate Eq. (39), (41) or (43). If the hyperparameters are to be estimated, an additional maximization step is included depending on the prior distribution $p(w_n)$. The iteration is repeated until convergence is achieved.

During each iteration we check the convergence with

$$L(q_k) - L(q_{k-1}) < \text{lTol}, \tag{70}$$

In our implementation the default value is $\text{lTol} = 10^{-8}$. We also monitor the convergence of the algorithm by comparing the absolute and relative change of the estimate of $\mathbf{x}$ in the iteration k to the estimate in the previous iteration $k-1$. The stopping criteria for the algorithm in the iteration k is

$$\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_\infty < \max\{\text{absTol}, \text{relTol} \cdot \|\mathbf{x}_k\|_\infty\}. \tag{71}$$

Our default values are $\text{abstol} = 10^{-8}$ and $\text{reltol} = 10^{-8}$. Also, the lower bound given in Eq. (60) can be used to monitor the convergence. The pseudocode for the algorithm is given in Algorithm 1.

The algorithm is implemented in the MATLAB function `rmvregress`, which also works in Octave. The default measurement noise distribution is the multivariate t with 4 degrees of freedom. The function can be freely downloaded from the Matlab Central file exchange.

Algorithm 1 VB algorithm for robust linear regression

Initialize: $\bar{\mathbf{x}} \leftarrow \underline{0}$, $\mathbf{S} \leftarrow \mathbf{I}$, $\mathbf{P} \leftarrow \mathbf{I}$, $\bar{\mathbf{w}} \leftarrow 1$

while **do**

for $n = 1$ **to** N **do**

Update $\bar{\mathbf{y}}_n$ using (31), (22), (32)

Update Σ_n using (28)

end for

$\mathbf{P} \leftarrow \left(\sum_{n=1}^{N} \bar{w}_n \mathbf{H}_n^T \mathbf{S} \mathbf{H}_n \right)^{-1}$

$\bar{\mathbf{x}} \leftarrow \mathbf{P} \left(\sum_{n=1}^{N} \bar{w}_n \mathbf{H}_n^T \mathbf{S} \bar{\mathbf{y}}_n \right)$

$\mathbf{R} \leftarrow \sum_{n=1}^{N} \bar{w}_n \left((\bar{\mathbf{y}}_n - \mathbf{H}_n \bar{\mathbf{x}})(\bar{\mathbf{y}}_n - \mathbf{H}_n \bar{\mathbf{x}})^T + \Sigma_n + \mathbf{H}_n \mathbf{P} \mathbf{H}_n^T \right)$

$\mathbf{S} \leftarrow (N + m)(\mathbf{A} + \mathbf{R})^{-1}$

for $n = 1$ **to** N **do**

Update $\bar{w}_n$ using (39), (41) or (43)

end for

if Hyperparameters unknown **then**

if $p(w_n)$ is gamma distribution **then**

Update α and β by maximizing (63) and (64)

else if $p(w_n)$ is inverse-gamma distribution **then**

Update α and β by maximizing (67) and (68)

else if $p(w_n)$ is contaminated normal distribution **then**

Update ϵ using (69)

end if

end if

Update lower bound using (60)

Check convergence using (70), (71)

end while

4 Examples

4.1 Stack Loss data, univariate observations

The stack loss dataset has been analyzed in the literature, for example Lange et al. (1989) and Hoeting et al. (1996). The dataset contains univariate observations of stack loss, which is assumed to depend linearly on three regressors: air flow, temperature, and acid content. Also an intercept term is estimated.

We consider robust regression with three different noise models: Student t, Laplace and contaminated normal distributions. We consider Student t distributions with degrees of freedom $\nu = 4$ and $\nu = 1.1$. The value $\nu = 4$ is a "general-purpose" choice, while $\nu = 1.1$ is the ML estimate for the degrees of freedom obtained in (Lange et al., 1989). The prior distribution $p(\mathbf{x}, Q)$ is taken to be the noninformative Jeffreys prior. The observed data is collected to a 21 element column vector Y. The predictor variables are collected into a 21×4 matrix H, where each row represents the values for the corresponding observation in Y. The first element of each row is a 1, corresponding to the intercept term.

Regression with Student's t-distribution with $\nu = 4$ degrees of freedom is the default option, so we use the command

```
>> [x,s,W]=rmvregress(H,Y);
```

For the degrees of freedom $\nu = 1.1$ use:

```
>> [x,s,W]=rmvregress(H,Y, 'student', 1.1);
```

The Laplace regression is obtained with

```
>> [x,s,W]=rmvregress(H,Y, 'laplace');
```

For the contaminated normal we take $\epsilon = 0.1$ and $c = 10$:

```
>> [x,s,W]=rmvregress(H,Y, 'contnorm', [0.1,10]);
```

The values of the lower bound during the VB updates are plotted in Fig. (1). As can be seen from the plot, the VB updates converges rather quickly close to the maximal value of the lower bound.

Small values in the estimated weights W correspond to possible outliers in the data. The weights for different observation distributions are collected to Table 1. As stated for example in (Lange et al., 1989) and (Hoeting et al., 1996) the observations 1,3,4 and 21 are generally considered outliers, so these rows are shaded in the table. We see that the Student t and Laplace distributions assign small weights to these four observations. The contaminated normal distribution clearly distinguishes the observations 4 and 2, but observations 1 and 3 do not stand out from the other observations.

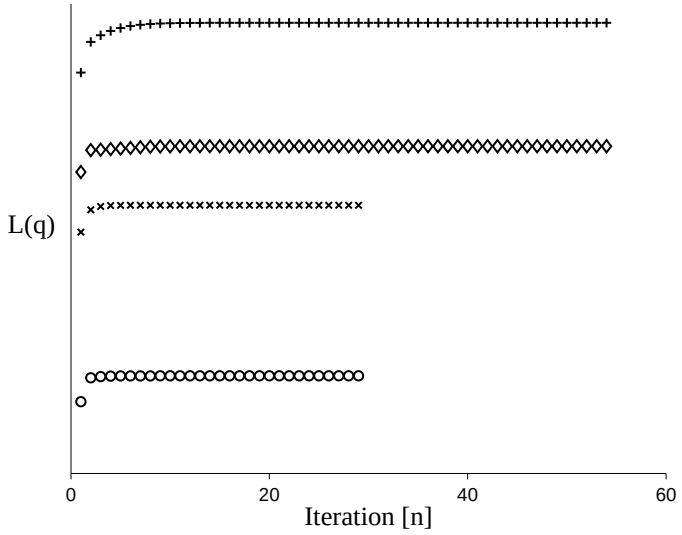

Figure 1: Computed values of the lower bound during the VB algorithm for t distribution ($nu = 4$) (circle), t distribution ($\nu = 1.1$) (plus), Laplace distribution (cross) and contaminated normal distribution (diamond).

Instead of just point estimates we may be interested in the full posterior distribution of the parameters. The covariance matrix for the regression coefficients is obtained by including P in the output argument list:

```
>> [x,s,W, P]=rmvregress(H,Y);
```

Table 1: Expected weights w_n for the observations

Observation	t_4	$t_{1.1}$	Laplace	Contam.
1	0.80	0.11	0.98	0.94
2	1.02	1.27	3.44	0.94
3	0.68	0.10	0.88	0.90
4	0.42	0.05	0.63	0.37
5	1.12	1.23	3.63	0.96
6	1.00	0.85	2.40	0.95
7	1.09	1.45	3.78	0.96
8	1.18	1.46	5.79	0.97
9	1.04	1.08	2.78	0.96
10	1.19	1.63	5.99	0.97
11	1.12	1.37	3.73	0.96
12	1.13	1.57	4.41	0.96
13	0.96	0.34	1.69	0.94
14	1.15	0.79	3.18	0.96
15	1.01	0.84	2.55	0.95
16	1.18	1.69	6.51	0.97
17	1.12	1.34	3.68	0.96
18	1.20	1.70	7.41	0.97
19	1.19	1.39	5.93	0.97
20	1.12	0.71	2.82	0.96
21	0.27	0.04	0.51	0.10

Table 2: Standard errors for the estimated coefficients. Asymptotic standard errors from Table 7 of (Lange et al., 1989).

	x_1	x_2	x_3	x_4
t_4	8.53	0.11	0.29	0.11
$t_{1.1}$	4.28	0.06	0.15	0.06
Laplace	5.97	0.08	0.21	0.08
Contam.	8.43	0.11	0.29	0.11
$t_{1.1}$ (asymptotic)	4.7	0.054	0.147	0.063

The standard errors for the regression coefficients can be obtained by taking the square root of the diagonal elements of P. The standard errors for the different regression models are listed in Table 2. We see that the standard errors for the $t_{1.1}$ distribution are close to the asymptotic results obtained with the expected information matrix in (Lange et al., 1989).

4.2 Astronomy data, bivariate observations

We use the star cluster data set of (Rousseeuw and Leroy, 2003) to illustrate robust estimation of the mean and the covariance matrix. The astronomy data is bivariate ($d = 2$), consisting of logarithms of the effective temperature at the surface of the star and of the light intensity of the star . We fit a multivariate t-distribution with $\nu = 5$ degrees of freedom, a multivariate Laplace distribution and a contaminated normal distribution to the data. The prior distribution $p(\mathbf{x}, \mathbf{Q})$ is taken to be the noninformative Jeffreys prior.

The data is collected in the 47×2 matrix $\mathbf{Y}$, where each row represents one observation. The design matrix is $\mathbf{H}_n = \mathbf{I}$ for all the observations. The regression is then performed by the command

```
>> [x,Q,W,P,R,V]=rmvregress(H,Y, 'student', 5);
```

The regression with the Laplace distribution is computed using the command

```
>> [x,Q,W,P,R,V]=rmvregress(H,Y, 'laplace');
```

For the contaminated normal distribution we take $\epsilon = 0.1$ and $c = 10$:

```
>> [x,Q,W,P,R,V]=rmvregress(H,Y, 'contnorm', [0.1, 10]);
```

The values of the lower bound during the VB updates are plotted in Fig. (2).

The obtained mean estimates for the location and the concentration ellipses, defined as (Anderson, 2003)

$$(\mathbf{x} - \mathbf{m})^T \mathbf{Q}^{-1} (\mathbf{x} - \mathbf{m}) = d + 2,$$

are plotted in Fig. (3). In the figure the extreme observations 7, 11, 20, 30 and 34 are labeled. The weights for these observations are listed in table 3. The rest of the weights are in the range $(0.55, 1.40)$ for the multivariate t, $(0.86, 25.50)$ for the Laplace and $(0.76, 0.99)$ for the contaminated normal. The Laplace distribution tends to heavily weight the observations closest to the mean. The contaminated normal distribution tends to assign very small weights to outliers and weights for non-outliers are all close to 1.

Fig. (3) shows also the mean and concentration ellipse obtained using ordinary least squares regression. It can be seen that the least squares regression results are significantly influenced by the extreme observations.

To gain some idea about the quality of the variational approximation, the computed means and 95% confidence intervals are compared to the results obtained using Gibbs sampler MCMC method. The Gibbs sampler is used to generate 5000 samples from the posterior distribution for each of the three families of robust regression distributions. The generated samples are used to obtain estimates for the mean and 95% confidence interval.

4.3 Risk research data, multivariate observations, missing data

In the third example we analyze St. Louis Risk Research data from (Little and Rubin, 2002, p. 119). The dataset contains information from 69 families with two children. The families are classified into three groups according to the mental health history of the parents. Group 1 is a control group of families, group 2 is a moderate risk group and group 3 is a high risk

Table 3: Expected weights w_n for the observations

Observation	t_5	Laplace	Contam.
7	0.37	0.69	0.17
11	0.12	0.35	0.10
20	0.12	0.34	0.10
30	0.11	0.33	0.10
34	0.10	0.32	0.10

Table 4: Estimated means and 95% confidence intervals for $\mathbf{x}$.

		x_1	95% CI	x_2	95% CI
t ($\nu = 5$)	VB	4.3937	(4.3478, 4.4338)	4.9591	(4.8062, 5.0782)
	MCMC	4.3934	(4.3414, 4.4414)	4.9604	(4.7969, 5.1182)
Laplace	VB	4.4056	(4.3718, 4.4395)	5.0296	(4.9309, 5.1283)
	MCMC	4.4067	(4.3582, 4.4546)	5.0362	(4.8617, 5.2054)
Cont. norm.	VB	4.3908	(4.3469, 4.4347)	4.9422	(4.7964, 5.0880)
	MCMC	4.3898	(4.3400, 4.4374)	4.9422	(4.7876, 5.0953)

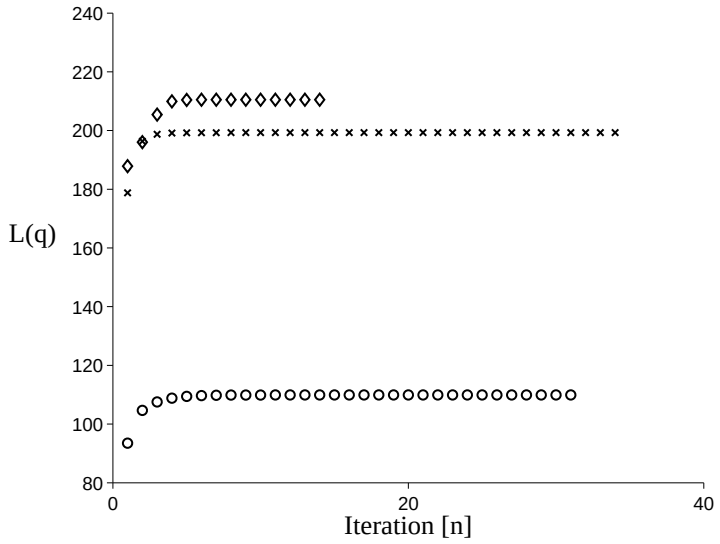

Figure 2: Computed values of the lower bound during the VB algorithm for t distribution (circle), Laplace distribution (cross) and contaminated normal distribution (diamond).

group. The data consists of standardized reading R and verbal V comprehension scores for both of the children in each family. For some children the R or V or both are missing.

We wish to study if the reading and verbal scores in group 1 are better than in the combined group of 2 and 3. Each observation consists of four scores $\mathbf{Y}_n = [R_1, V_1, R_2, V_2]^T$, which are the reading and verbal comprehension scores for the first and second child respectively. The missing values are assigned an value NaN. We form a regression model with eight regression parameters

$$\mathbf{Y}_n = \left[\begin{array}{cc} \mathbf{H}_1 & \mathbf{H}_2 \end{array} \right] \mathbf{x} + \mathbf{v},$$

where $\mathbf{H}_k = \mathbf{I}$, if the observation n was from group k and $\mathbf{H}_k = 0$ otherwise. The regression coefficients are

$$\mathbf{x} = \left[\begin{array}{cc} \mathbf{x}_1^T & \mathbf{x}_2^T \end{array} \right]^T = \left[\begin{array}{cccccccc} R_1^{(1)} & V_1^{(1)} & R_2^{(1)} & V_2^{(1)} & R_1^{(2)} & V_1^{(2)} & R_2^{(2)} & V_2^{(2)} \end{array} \right]^T.$$

For the measurements we consider multivariate t with 4 degrees of freedom, multivariate Laplace and contaminated normal distribution with $\epsilon = 0.1$ and $c = 10$. We use the commands:

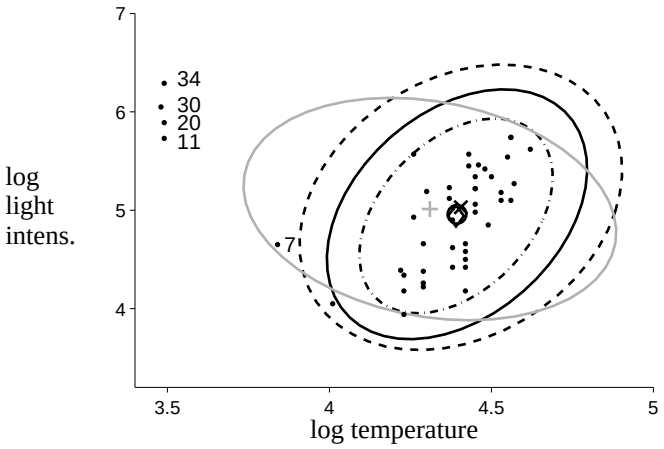

Figure 3: Fitted means and concentration ellipses for Student t distribution (circle, solid), Laplace distribution (cross, dashed), contaminated normal distribution (diamond, dash-dot) and least squares regression (gray plus, gray solid).

```
>> [m,Q,W,P,R,V]=rmvregress(H,Y);

>> [m,Q,W,P,R,V]=rmvregress(H,Y, 'laplace');

>> [m,Q,W,P,R,V]=rmvregress(H,Y, 'contnorm', [0.1, 10]);
```

The values of the lower bound during the VB updates are plotted in Fig. (4). We are interested in the difference in the means of the two groups, i.e. in the posterior distribution of $\mathbf{x}_1 - \mathbf{x}_2$. The joint posterior for $\mathbf{x} = \left[\begin{smallmatrix} \mathbf{x}1 \\ \mathbf{x}_2 \end{smallmatrix}\right]$ is normal with mean $\mathbf{m}$ and variance $\mathbf{P}$. From this it follows that $\mathbf{x}_1 - \mathbf{x}_2$ has a normal distribution with mean $\mathbf{m}_1 - \mathbf{m}_2$ and variance $\mathbf{P}_1 + \mathbf{P}_2 - \mathbf{P}_{1,2} - \mathbf{P}_{1,2}^T$. The marginal posteriors for each of the four differences are plotted in Fig. 5. The 95% posterior probability intervals are also listed in table 5. The conclusions using all the distributions are that the scores are better in the group 1. We see that our results are consistent with the results in (Little and Rubin, 2002, p. 261), where the posterior histograms and 95% probability intervals were obtained using Monte Carlo methods for multivariate t distribution. As can be seen from the Monte Carlo results using the multivariate t distribution, the variational approximation for the posterior tends to underestimate the true posterior variance.

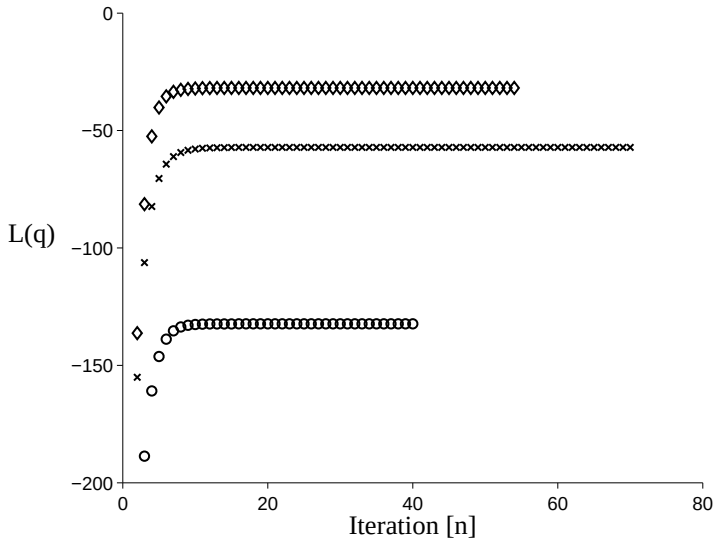

Figure 4: Computed values of the lower bound during the VB algorithm for t distribution ($nu = 4$) (circle), Laplace distribution (cross) and contaminated normal distribution (diamond).

5 Conclusion

In this paper we considered multivariate robust linear regression with missing data. We concentrated on robust models in which the observations can be written as a Gaussian scale mixtures. Specifically we considered three families of distributions corresponding to different priors for the weights in the Gaussian scale mixture presentation. These families include as special cases the multivariate t, multivariate Laplace, and multivariate contaminated normal models that are often used for robust statistical modeling.

Robust linear regression is much studied in the statistical literature. Algorithms for statistical inference are mainly based on the EM algorithm or on Monte Carlo methods. In this paper we presented a variational Bayes method to compute an approximation for the posterior distribution. The VB method can be seen as an alternative to the computationally heavy Monte Carlo methods in Bayesian inference. The VB algorithm is computationally comparable to the EM algorithm, but is better suited for full Bayesian inference because we obtain directly the full approximate posterior for the model parameters. The VB method can

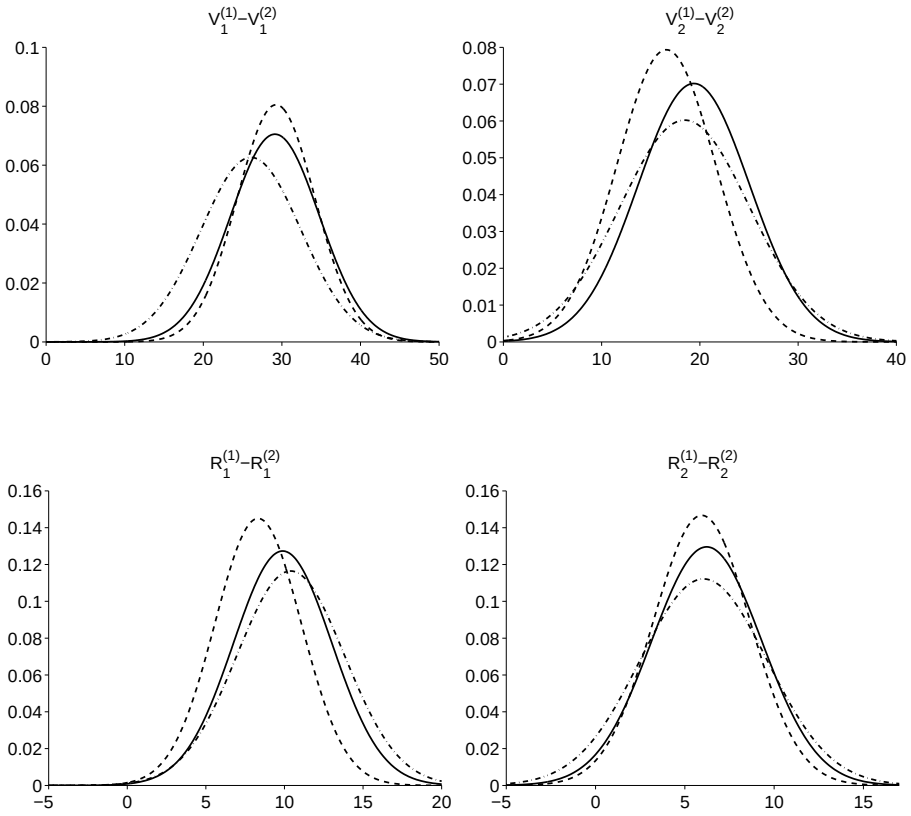

Figure 5: Marginal posterior distributions for the difference $\mathbf{x}_1 - \mathbf{x}_2$. Student t-distribution (solid), Laplace distribution (dashed) and contaminated normal distribution (dash-dot).

be also extended to include the estimation of the hyperparameters for the prior distribution of weights in the Gaussian scale mixture presentation.

Examples from literature were presented to illustrate the use of our MATLAB/Octave implementation and to compare the results to those obtained in the references. The implementation can be freely downloaded from the Matlab Central file exchange.

The proposed method for variational inference is based on fully factorizing the posterior distribution, which also provides directly the posterior marginal distributions for all the variables. The accuracy of the approximation is in general difficult to access, however the variance of the approximated distribution tends to underestimate the true variance of the posterior (MacKay, 2003). This result is also observed for the examples 2 and 3, where the confidence intervals obtained using MCMC methods tend to be wider than confidence

Table 5: 95% posterior probability intervals for the difference $\mathbf{x}_1 - \mathbf{x}_2$. MCMC results from (Little and Rubin, 2002, p. 261)

.

	t_4 (MCMC)	t_4	Laplace	Contam.
$V_1^{(1)} - V_1^{(2)}$	$(15.38, 43.99)$	$(18.04, 40.21)$	$(19.56, 38.98)$	$(13.58, 38.55)$
$V_2^{(1)} - V_2^{(2)}$	$(7.08, 31.42)$	$(8.25, 30.53)$	$(6.73, 26.44)$	$(5.47, 31.44)$
$R_1^{(1)} - R_1^{(2)}$	$(1.68, 17.94)$	$(3.75, 16.03)$	$(2.93, 13.72)$	$(3.69, 17.11)$
$R_2^{(1)} - R_2^{(2)}$	$(-0.60, 14.04)$	$(0.19, 12.26)$	$(0.62, 11.28)$	$(-0.92, 13.02)$

intervals obtained using the variational method.

In this paper the lower bound was used mainly for inspecting the convergence of the VB algorithm. However, this can be also utilized for model comparison as discussed for example in (Beal, 2003, pp. 60-61). Also, differences of the VB method with the EM algorithm could be studied by comparing values of the lower bound with the values of log-likelihood during the iterations.

Interesting extension for the robust methods considered in this paper would be to apply them for nonlinear problems and time-series analysis. Variational method for nonlinear forward models using Gaussian noise is considered for example in (Chappell et al., 2009). Combining these methods to deal with nonlinear models with the Gaussian scale mixture presentation of the measurement distribution could be used to obtain variational methods for the nonlinear robust regression. The variational inference has also recently been applied for robustifying the Kalman filter algorithm using t distributed noise for the measurements (Agamennoni et al., 2011; Piché et al., 2012). Kalman filters with more general measurement distributions could be obtained by utilizing the Gaussian scale mixture models considered in this paper.

Acknowledgments

The first author received financial support from the Tampere Doctoral Programme in Information Science and Engineering (TISE).

References

Abramowitz, M. and Stegun, I. (1965). *Handbook of Mathematical Functions*. Dover Publications.

Agamennoni, G., Nieto, J. I., and Nebot, E. M. (2011). An outlier-robust kalman filter. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1551–1558. IEEE.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley.

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–102.

Beal, M. J. (2003). *Variational Algorithms for approximate Bayesian Inference*. PhD thesis, University of London.

Beal, M. J. and Ghahramani, Z. (2003). The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian Statistics 7*, pages 453–463. Oxford University Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bishop, C. M., Winn, J., and Spiegelhalter, D. (2002). Vibes: A variational inference engine for bayesian networks. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems XV*.

Blattberg, R. C. and Gonedes, N. J. (1974). A comparison of the stable and student distributions as statistical models for stock prices. *The Journal of Business*, 47(2):pp. 244–280.

Bloomfield, P. and Steiger, W. L. (1983). *Least absolute deviations : theory, applications, and algorithms*. Progress in probability and statistics. Birkhäuser.

Chappell, M., Groves, A., Whitcher, B., and Woolrich, M. (2009). Variational bayesian inference for a nonlinear forward model. *Signal Processing, IEEE Transactions on*, 57(1):223 –236.

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

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis, 2nd edition*. Chapman & Hall/CRC.

Hoeting, J., Hoeting, J., Raftery, A. E., and Madigan, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *COMPUTATIONAL STATISTICS AND DATA ANALYSIS*, 22:25–70.

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

Jørgensen, B. (1982). *Statistical properties of generalized inverse gaussian distribution*. Lecture Notes in Statistics. Springer.

Kotz, S., Kozubowski, T. J., and Podgórski, K. (2001). *The Laplace distribution and generalizations*. Birkhäuser.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t distributions and their applications*. Cambridge University Press.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):pp. 881–896.

Little, R. J. A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 37(1):pp. 23–38.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley.

Liu, C. (1996). Bayesian robust multivariate linear regression with incomplete data. *Journal of the American Statistical Association*, 91(435):1219–1227.

Liu, C. and Rubin, D. B. (1994). The ecme algorithm: A simple extension of em and ecm with faster monotone convergence. *Biometrika*, 81(4):pp. 633–648.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter expansion to accelerate em: The px-em algorithm. *Biometrika*, 85(4):pp. 755–770.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. John Wiley & Sons, 2nd edition edition.

Meng, X.-L. and Dyk, D. v. (1997). The em algorithm–an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3):pp. 511–567.

Meng, X.-L. and Rubin, D. (1993). Maximum likelihood estimation via the ecm algorithm: a general framework. *Biometrika*, 80(2):267–278.

Meng, X.-L. and Rubin, D. B. (1991). Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86(416):pp. 899–909.

Ormerod, J. and Wand, M. (2010). Explaining variational approximations. *The American Statistician*, 64(2).

Penny, W., Kilner, J., and Blankenburg, F. (2007). Robust bayesian linear models. *Neuroimage*, 36(3):661–671.

Phillips, R. (2002). Least absolute deviations estimation via the em algorithm. *Statistics and Computing*, 12:281–285.

Piché, R., Sarkka, S., and Hartikainen, J. (2012). Recursive outlier-robust filtering and smoothing for nonlinear systems using the multivariate student-t distribution. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE.

Roberts, S. and Penny, W. (2002). Variational bayes for generalized autoregressive models. *Signal Processing, IEEE Transactions on*, 50(9):2245 – 2257.

Rousseeuw, P. J. and Leroy, A. M. (2003). *Robust Regression and Outlier Detection*. Wiley.

Rubin, D. B. (2004). Iteratively reweighted least squares. In *Encyclopedia os Statistical Sciences*. John Wiley & Sons.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):pp. 528–540.

Tipping, M. and Lawrence, N. (2003). A variational approach to robust bayesian interpolation. In *Neural Networks for Signal Processing, 2003. NNSP'03. 2003 IEEE 13th Workshop on*, pages 229 – 238.

Titterington, D. (2004). Bayesian methods for neural networks and related models. *Statistical Science*, 19(1):128–139.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 448–485. Stanford University Press.

Verboven, S. and Hubert, M. (2005). Libra: a matlab library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, 75(2):127 – 136.

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

Wand, M. P., Ormerod, J. T., Padoan, S. A., and Frührwirth, R. (2011). Mean field variational bayes for elaborate distributions. *Bayesian Analysis*, 6(4):1–48.

West, M. (1984). Outlier models and prior distributions in bayesian linear regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):pp. 431–439.

Zellner, A. (1976). Bayesian and non-bayesian analysis of the regression model with multivariate student-t error terms. *Journal of the American Statistical Association*, 71(354):pp. 400–405.