Robust multivariate linear regression using the Student-t distribution

Robert Piché
Tampere University of Technology

***Abstract*— This document presents the theoretical background of the Matlab program `mvsregress` for multivariate linear regression based on the Student-t distribution.**

I. Introduction

The Matlab Statistics Toolbox has functions for robust univariate linear regression, and has a function `mvregress` for Normal multivariate linear regression, but has no functions for robust multivariate linear regression. To fill this gap, I have written a Matlab function `mvsregress` for multivariate linear regression based on the Student-t distribution. Because this distribution has "fat tails" compared to the Normal distribution, regression is robust, in the sense that it is less sensitive to extreme observations.

This document describes the statistical model and the numerical algorithm used in `mvsregress` for computing the maximum a-posteriori estimate of the model parameters. Two examples are presented to illustrate the robustness of Student regression compared to Normal regression.

II. Statistical model

Each d-variate observation $\mathbf{y}_n$ is modelled as a linear function of a k-variate parameter vector $\mathbf{x}$ with additive noise,

$$\mathbf{y}_n = \mathbf{H}_n\mathbf{x} + \text{noise}$$

Assuming the noise to be a zero-mean multivariate Student-t with shape matrix $\mathbf{Q}$ and ν degrees of freedom, the observation has the distribution

$$\mathbf{y}_n \,|\, \mathbf{x}, \mathbf{Q} \sim \text{Student}(\mathbf{H}_n\mathbf{x}, \mathbf{Q}, \nu) \tag{1}$$
$$p(\mathbf{y}_n \,|\, \mathbf{x}, \mathbf{Q}) \propto |\mathbf{Q}|^{\frac{1}{2}} \big(1 + \tfrac{1}{\nu}(\mathbf{y}_n - \mathbf{H}_n\mathbf{x})'\mathbf{Q}(\mathbf{y}_n - \mathbf{H}_n\mathbf{x})\big)^{-\frac{\nu+d}{2}}$$

This distribution can be obtained as a mixture of Normals (Fig. 1) by marginalising an auxiliary weight parameter w_n having the prior distribution

$$w_n \sim \text{Gamma}(\tfrac{\nu}{2}, \tfrac{\nu}{2}), \quad p(w_n) \propto w_n^{\frac{\nu}{2}-1} \mathrm{e}^{-\frac{\nu}{2}w_n}$$

out of

$$\mathbf{y}_n \,|\, \mathbf{x}, \mathbf{Q}, w_n \sim \text{Normal}(\mathbf{H}_n\mathbf{x}, (w_n\mathbf{Q})^{-1})$$
$$p(\mathbf{y}_n \,|\, \mathbf{x}, \mathbf{Q}, w_n) \propto |w_n\mathbf{Q}|^{\frac{1}{2}} \mathrm{e}^{-\frac{w_n}{2}(\mathbf{y}_n - \mathbf{H}_n\mathbf{x})'\mathbf{Q}(\mathbf{y}_n - \mathbf{H}_n\mathbf{x})}$$

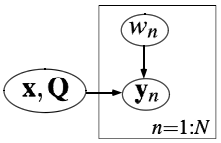

Fig. 1. Directed acyclic graph representation of the Student data model as a mixture of Normals

Assuming that N observations are conditionally independent, the likelihood density for the $d \times N$ observation array $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ is

$$p(\mathbf{Y} \,|\, \mathbf{x}, \mathbf{Q}, \mathbf{w}) = \prod_{n=1}^{N} p(\mathbf{y}_n \,|\, \mathbf{x}, \mathbf{Q}, w_n)$$
$$\propto |\mathbf{Q}|^{\frac{N}{2}} \cdot \prod_{n=1}^{N} w_n^{d/2} \cdot \mathrm{e}^{-\frac{1}{2}\sum_{n=1}^{N} w_n(\mathbf{y}_n - \mathbf{H}_n\mathbf{x})'\mathbf{Q}(\mathbf{y}_n - \mathbf{H}_n\mathbf{x})}$$

Assuming that $w_1, \ldots, w_N, \mathbf{x}, \mathbf{Q}$ are a-priori jointly independent, with the uninformative improper prior distribution

$$p(\mathbf{x}, \mathbf{Q}) \propto |\mathbf{Q}|^{-(d+1)/2}$$

leads to the posterior density

$$\begin{aligned} &p(\mathbf{x}, \mathbf{Q}, \mathbf{w} \,|\, \mathbf{Y}) \\ &\quad \propto p(\mathbf{Y} \,|\, \mathbf{x}, \mathbf{Q}, \mathbf{w}) p(\mathbf{x}, \mathbf{Q}) p(\mathbf{w}) \\ &\quad \propto |\mathbf{Q}|^{\frac{N-d-1}{2}} \cdot \prod_{n=1}^{N} w_n^{\frac{d+\nu}{2}-1} \\ &\qquad \cdot \mathrm{e}^{-\frac{1}{2}\sum_{n=1}^{N} w_n\left((\mathbf{y}_n - \mathbf{H}_n\mathbf{x})'\mathbf{Q}(\mathbf{y}_n - \mathbf{H}_n\mathbf{x}) + \nu\right)} \end{aligned} \tag{2}$$

By examination of the posterior (2), it can be seen that the posterior conditional weights are independently Gamma distributed:

$$w_n \,|\, \mathbf{Y}, \mathbf{x}, \mathbf{Q} \sim \text{Gamma}\Big(\frac{d+\nu}{2}, \frac{\nu + (\mathbf{y}_n - \mathbf{H}\mathbf{x})'\mathbf{Q}(\mathbf{y}_n - \mathbf{H}\mathbf{x})}{2}\Big)$$
$$p(\mathbf{w} \,|\, \mathbf{Y}, \mathbf{x}, \mathbf{Q}) = \prod_{n=1}^{N} p(w_n \,|\, \mathbf{Y}, \mathbf{x}, \mathbf{Q})$$

with posterior conditional means

$$\mathsf{E}(w_n \,|\, \mathbf{Y}, \mathbf{x}, \mathbf{Q}) = \frac{d+\nu}{\nu + (\mathbf{y}_n - \mathbf{H}\mathbf{x})'\mathbf{Q}(\mathbf{y}_n - \mathbf{H}\mathbf{x})} \tag{3}$$

The shape matrix's posterior conditional distribution is derived as follows. The quadratic form in (2) can be written as

$$\sum_{n=1}^{N} w_n(\mathbf{y}_n - \mathbf{H}_n\mathbf{x})'\mathbf{Q}(\mathbf{y}_n - \mathbf{H}_n\mathbf{x}) = \mathrm{tr}(\mathbf{QS})$$

where

$$\mathbf{S} = \sum_{n=1}^{N} w_n(\mathbf{y}_n - \mathbf{H}_n\mathbf{x})(\mathbf{y}_n - \mathbf{H}_n\mathbf{x})'$$

Using this fact, and by examination of the posterior (2), it can be seen that the posterior conditional distribution is

$$\mathbf{Q} \,|\, \mathbf{Y}, \mathbf{x}, \mathbf{w} \sim \mathrm{Wishart}(\mathbf{S}^{-1}, N)$$
$$p(\mathbf{Q} \,|\, \mathbf{Y}, \mathbf{x}, \mathbf{w}) \propto |\mathbf{Q}|^{\frac{N-d-1}{2}} \cdot \mathrm{e}^{-\frac{1}{2}\,\mathrm{tr}\,\mathbf{QS}}$$

Its mode is

$$\mathsf{mode}(\mathbf{Q} \,|\, \mathbf{Y}, \mathbf{x}, \mathbf{w}) = (N - d - 1)\mathbf{S}^{-1} \tag{4}$$

for $N \geq d + 1$.

The parameter vector's posterior conditional distribution is Normal, as can be seen by examination of the posterior (2). Its mode is

$$\mathsf{mode}(\mathbf{x} \,|\, \mathbf{Y}, \mathbf{Q}, \mathbf{w}) = \left(\sum_{n=1}^{N} w_n\mathbf{H}_n'\mathbf{Q}\mathbf{H}_n\right)^{-1} \sum_{n=1}^{N} w_n\mathbf{H}_n'\mathbf{Q}\mathbf{y}_n \tag{5}$$

III. ECM ALGORITHM

The elements of the Maximum A-Posteriori (MAP) estimate $\mathsf{mode}(\mathbf{x}, \mathbf{Q} \,|\, \mathbf{Y})$ can be computed using an Expectation Conditional Maximisation (ECM) algorithm [1], [2]. For the Student data model presented in the previous section, the ECM algorithm's E-step (expectation of the auxiliary parameters) uses (3), and the CM-steps (conditional maximisations) use (4) and (5).

The algorithm is

1. initialize $\mathbf{w} \leftarrow \mathbf{1}$ and $\mathbf{Q} \leftarrow \mathbf{I}$
2. for t from 1 to T do
3. $\quad \mathbf{x} \leftarrow \left(\sum_{n=1}^{N} w_n\mathbf{H}_n'\mathbf{Q}\mathbf{H}_n\right)^{-1} \sum_{n=1}^{N} w_n\mathbf{H}_n'\mathbf{Q}\mathbf{y}_n$
4. $\quad \mathbf{S} \leftarrow \sum_{n=1}^{N} w_n(\mathbf{y}_n - \mathbf{H}_n\mathbf{x})(\mathbf{y}_n - \mathbf{H}_n\mathbf{x})'$
5. $\quad \mathbf{Q} \leftarrow (N - d - 1)\mathbf{S}^{-1}$
6. $\quad$ for n from 1 to N do
7. $\quad\quad w_n \leftarrow \frac{d+\nu}{\nu + (\mathbf{y}_n - \mathbf{H}\mathbf{x})'\mathbf{Q}(\mathbf{y}_n - \mathbf{H}\mathbf{x})}$
8. $\quad$ end do
9. end do

The MAP estimate for Normal regression, which is the limiting case of Student regression with $\nu \to \infty$, can be obtained by omitting the weight update in lines 6–8.

The Matlab function `mvsregress` implements the above algorithm. The calling syntax is similar to that of the Matlab Statistics Toolbox function `mvregress`, as follows.

`[x,Q,w]=mvsregress(H,Y)` performs multivariate Student regression of the N multivariate d-variate observations in the $N \times d$ matrix `Y` on the predictor variables in `H`, and returns a k-element column vector `x` of MAP estimates of the regression coefficients $\mathbf{x}$, a $d \times d$ matrix `Q` of the MAP estimate of the Student scale matrix $\mathbf{Q}$, and an N-element vector `w` of the weights $\mathbf{w}$.

`H` may be either a matrix or a cell array. If $d = 1$, `H` may be an $N \times k$ design matrix of predictor variables. For any value of d, `H` may also be a cell array of length N, each cell containing the $d \times k$ design matrix $\mathbf{H}_n$ for one multivariate observation. If all observations have the same $d \times k$ design matrix, `H` may be a single cell.

`[x,Q,w]=mvsregress(H,Y,nu)` uses a Student distribution with `nu` degrees of freedom; the default is `nu` $= 5$. If `nu` is `inf` then Normal regression is done.

In `mvsregress`, the ECM iteration loop (lines 2–9 of the algorithm) is stopped if the change in $\mathbf{x}$ is small; at most $T = 100$ ECM iterations are made. There is no provision for missing values as in `mvregress`.

IV. EXAMPLES

A. Synthetic data, univariate observations

The Matlab Statistics Toolbox program `robustdemo` uses the data set

$$\mathbf{Y} = [-0.6867,\ 1.7258,\ 1.9117,\ 6.1832,\ 5.3636,\ 7.1139,\ 9.5668,\ 10.0593,\ 11.4044,\ 6.1677]$$

to demonstrate robust fitting of a straight line. Fitting the model (1) with $\mathbf{H}_n = [1\ n]$, the Student regression's MAP estimate for $\nu = 5$ degrees of freedom is

$$\hat{\mathbf{x}} = [-1.2657,\ 1.3828]'$$

The corresponding straight line fit can be seen (Fig. 2) to be less affected by the outlying 10th observation than a Normal (least-squares) fit.

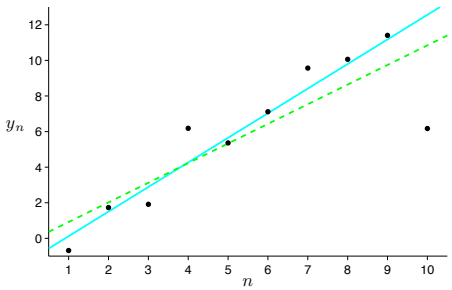

Fig. 2. A scatterplot of synthetic data, fitted Student line (solid), and fitted Normal line (dashed).

B. Astronomy data, bivariate observations

A set of 47 bivariate astonomical observations is used in [3] to illustrate robust regression. A Student distribution can be fitted to the data by using the model (1) with $\mathbf{H} = \mathbf{I}$. With $\nu = 5$ degrees of freedom, the Student regression MAP estimate $(\hat{\mathbf{x}}, \hat{\mathbf{Q}})$ is

$$\left(\begin{bmatrix} 4.3919 \\ 4.9588 \end{bmatrix}, \begin{bmatrix} 44.3028 & -4.8917 \\ -4.8917 & 4.6122 \end{bmatrix}\right)$$

The Normal regression MAP estimate is

$$\left(\begin{bmatrix} 4.3100 \\ 5.0121 \end{bmatrix}, \begin{bmatrix} 11.8332 & 1.2676 \\ 1.2676 & 3.0670 \end{bmatrix} \right)$$

The MAP estimates' parameters and covariance ellipses

$$\frac{\nu}{\nu - 2}(\mathbf{x} - \hat{\mathbf{x}})' \hat{\mathbf{Q}} (\mathbf{x} - \hat{\mathbf{x}}) = 1$$

are shown in Figure 3. It can be seen that the Student distribution is better aligned with the main cluster of points than the Normal, which is strongly influenced by the four points in the northwest corner.

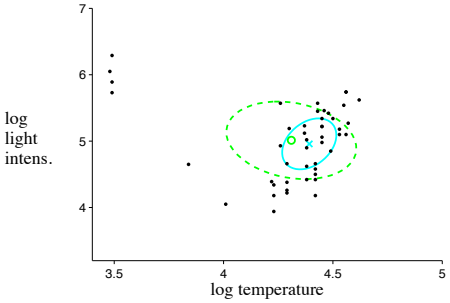

Fig. 3. An astronomy data scatterplot, fitted Student distribution's mean (x) and covariance ellipse (solid line), and fitted Normal distribution's mean (o) and covariance ellipse (dashed line).

REFERENCES

[1] X. L. Meng and D. B. Rubin, Maximum likelihood via the ECM algorithm: a general framework, *Biometrika*, **80**, 267–278, 1993.

[2] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed., Wiley, 2002.

[3] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, Wiley, 2003.