Computational Aspects on S- and CM-estimation

H. Ekblom[1], and O. Edlund[1]

[1] Luleå University of Technology, Department of Mathematics, S-97187 Luleå, Sweden

Keywords: CM-estimators, S-estimators, Robust estimation, Robust regression, Algorithms.

Robust estimation methods for regression have been developed for many years. Well-known examples are M-estimates (Huber, 1981) and S-estimates (Rousseeuw and Yohai, 1984). A more recent approach is "Constrained M-estimates", or CM-estimates for short, which has attractive statistical properties (Mendes and Tyler, 1995). We can formulate CM-estimation for regression the following way.

Consider the linear model

$$y_i = x_i^T \beta + e_i, \quad i = 1, 2, \ldots, n$$

where $y = (y_1, y_2, \ldots, y_n)^T$ is the response vector, x_i is the i:th row in the $(n \times p)$ design matrix X, β a p-dimensional vector of unknowns and $e = (e_1, e_2, \ldots, e_n)^T$ the error vector.

Define the residuals as $r_i = y_i - x_i^T \beta,\ i = 1, 2, \ldots, n$. Using the notation "ave" for the arithmetic average, the CM-estimation problem is to find the global minimum of

$$L(\beta, \sigma) = \text{ave}\{\rho_c(r_i/\sigma)\} + \log(\sigma)$$

over $\beta \in \underline{\mathrm{R}}^p$ and $\sigma \in \underline{\mathrm{R}}_+$ subject to the constraint

$$\text{ave}\{\rho_c(r_i/\sigma)\} \leq \varepsilon \rho_c(\infty) \tag{1}$$

Here $\rho_c(t)$ is a bounded, nondecreasing function of $t \geq 0$ with tuning parameter $c > 0$. If strict inequality holds in the constraint (1) we get the redescending M-estimating equations for β and σ. To find the S-estimate, we minimize L with respect to σ. This implies equality in the constraint (1). The CM-estimates, which are the solutions of a nonlinear minimization problem with an inequality constraint, cannot be expressed explicitly. Computing CM-estimates numerically is a challenging problem, since we like to minimize an object function where many local minima exist. In Arslan et al. (2001) an algorithm is presented for linear S- and CM-regression. We have modified this code e.g. to register all local minima of the object function. By running the code extensively we can get an answer to the question how many local minima do exist. We have performed such investigations for real as well as artificial problems. On these problems we can also see how well different algorithms manage to find the global minimum within a given time frame. We will present the result of such an investigation.

The corresponding multivariate estimation problem can be formulated the following way. Let $\mathrm{X}_n = \{x_1, x_2, \ldots, x_n\}$ be a data set in $\underline{\mathrm{R}}^p$, $p \geq 1$, and consider the problem of estimating the location and scatter parameters of X_n. CM-estimates for multivariate estimation, which have good local and global robustness properties, were introduced by Kent and Tyler (1996). They are defined as the global minimum of the objective function

$$L(\mu, \Sigma; X_n) = \text{ave}\{\rho(s_i)\} + \frac{1}{2} \log |\Sigma|$$

over $\mu \in \underline{\mathrm{R}}^p$ and $\Sigma \in P$ subject to the constraint

$$\text{ave}\{\rho(s_i)\} \leq \varepsilon \rho(\infty) \tag{2}$$

where P is the set of positive definite symmetric matrices, $s_i = (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$, for $i = 1, \ldots, n$, $0 < \varepsilon < 1$, and $\rho(s)$ is bounded, nondecreasing function of $s \geq 0$. Here, μ and Σ are unknown location and scatter parameters.

Similar to the regression case, when constraint (2) reduces to an equality, the CM-estimates will be the S-estimates for the location and the scatter parameters of the data (Lopuhaä, 1989).

We will discuss different algorithmic approaches and also explain why the regression code (Arslan et al., 2001) cannot be immediately generalized to the multivariate case.

What *may* be possible, on the other hand, is to generalize the method for finding M-estimates in regression models with *non-linear* dependence on β (Edlund et al., 1997), to also handle S- and CM-estimates. This may be accomplished by using techniques similar to, though modified from, the ones presented here and in Arslan et al. (2001).

References

O. Arslan, O. Edlund and H. Ekblom (2001). Algorithms to compute CM- and S-estimates for regression. Tech. report, Dept of Mathematics, Luleå University of Technology, Sweden (to be published in Metrica).

O. Edlund (1997). Linear M-estimation with bounded variables. *BIT*, 37, 13–23.

O. Edlund, H. Ekblom and K. Madsen (1997). Algorithms for non-linear M-estimation. *Computational Statistics*, 12, 373–383.

P.J. Huber (1981). *Robust Statistics*. Wiley, New York.

J.T. Kent and D.E. Tyler (1996). Constrained M-estimation for multivariate location and scatter. *The Annals of Statistics*, 24(3), 1346–1370.

H.P. Lopuhaä (1989). On the relationship between S-estimators and M-estimators of multivariate location and covariance. *The Annals of Statistics*, 17, 1662–1684.

B. Mendes and D.E. Tyler (1995). Constrained M estimates for regression. In: *Robust Statistics; Data Analysis, and Computer Intensive Methods. Lecture Notes In Statistics 109*, pp. 299–320. Springer, New York.

P.J. Rousseeuw and V.J. Yohai (1984). Robust regression by means of S-estimators. In: J. Frank, W. Härdle and R.D. Martin, editors, *Robust and Nonlinear Time Series Analysis*, pp. 256–272. Springer-Verlag, New York.

D. Ruppert (1992). Computing S estimators for regression and multivariatelocation/dispersion. *Journal of Computational ad Graphical Statistics*, 1, 253–270.