Robust Inference

October 6, 2015

Robust Multivariate Location and Scatter Estimates

Some Background

We begin by stating an interesting result.

Lemma 1 *(Grubel, 1988) Let* $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ *be a sample of* p *dimensional vectors in* R^p *with sample mean* $\overline{\mathbf{x}}$ *and sample covariance*

$$S = \frac{1}{n} \sum^{n} (\mathbf{x}_i - \overline{\mathbf{x}}) (\mathbf{x}_i - \overline{\mathbf{x}})'.$$

Consider the set of all pairs $(A, \mathbf{t})$ *such that* A *is positive definite and*

$$\frac{1}{np} \sum^{n} (\mathbf{x}_i - \mathbf{t})' A^{-1} (\mathbf{x}_i - \mathbf{t}) = 1. \tag{1}$$

Then the pair $(S, \overline{\mathbf{x}})$ *satisfies (1) and* S *minimizes the* $\det(A)$, *subject to (1).*

Proof of Lemma 1

First we show that $(S, \overline{\mathbf{x}})$ satisfies constraint (1):

$$\begin{aligned}
\frac{1}{np}\sum^{n}\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)' S^{-1}\left(\mathbf{x}_i-\overline{\mathbf{x}}\right) &= \frac{1}{p}\frac{1}{n}\left(\sum^{n} tr\left[\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)' S^{-1}\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)\right]\right) \\
&= \frac{1}{p}\left(\frac{1}{n}\sum^{n} tr\left[\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)' S^{-1}\right]\right) \\
&= \frac{1}{p}\left(tr\left(\frac{1}{n}\sum^{n}\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)' S^{-1}\right)\right) \\
&= \frac{1}{p}\left(tr\left(SS^{-1}\right)\right)=\frac{p}{p}=1
\end{aligned}$$

Suppose, now, that

$$\frac{1}{np}\sum^{n}\left(\mathbf{x}_i-\mathbf{t}\right)' A^{-1}\left(\mathbf{x}_i-\mathbf{t}\right)=1.$$

It is easy to show that there exist a pair $(B, \overline{\mathbf{x}})$ with $\det(B) \leq \det(A)$. Hence, it suffices to prove that

$$\det(B) \geq \det(S)$$

subject to

$$\bar{\mathbf{y}}=\mathbf{0} \quad \text{and} \quad \frac{1}{np}\sum^{n}\mathbf{y}_i' B^{-1}\mathbf{y}_i=1.$$

Moreover, since

$$\begin{aligned}
\frac{1}{np}\sum^{n}\mathbf{y}_i' B^{-1}\mathbf{y}_i &= \frac{1}{np}\sum^{n}\mathbf{y}_i' S^{-1/2}S^{1/2}B^{-1}S^{1/2}S^{-1/2}\mathbf{y}_i \\
&= \frac{1}{np}\sum^{n}\left(S^{-1/2}\mathbf{y}_i\right)'\left[S^{1/2}B^{-1}S^{1/2}\right]\left(S^{-1/2}\mathbf{y}_i\right) \\
&= \frac{1}{np}\sum^{n}\mathbf{z}_i' C^{-1}\mathbf{z}_i
\end{aligned}$$

$$\mathbf{z}_i=S^{-1/2}\left(\mathbf{x}_i-\overline{\mathbf{x}}\right), \quad \bar{\mathbf{z}}=\mathbf{0}, \quad S_{\mathbf{z}}=I$$

It suffices to prove that

$$\det(C)>1$$

subject to

$$\bar{\mathbf{z}}=\mathbf{0} \quad , \quad S_{\mathbf{z}}=I \quad \text{and} \quad \frac{1}{np}\sum^{n}\mathbf{z}_i' C^{-1}\mathbf{z}_i=1$$

with equality if and only if $C = I$.

Since C is symmetric and positive definite, there exists an orthogonal matrix P and a diagonal matrix $D = diag\left(\lambda_1, \lambda_2, ..., \lambda_p\right)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ such that

$$\begin{aligned} C &= P'DP \\ C^{-1} &= P'D^{-1}P \\ \det\left(C\right) &= \lambda_1\lambda_2...\lambda_p \end{aligned}$$

Now,

$$\begin{aligned} 1 &= \frac{1}{np}\sum^{n}\mathbf{z}_i'C^{-1}\mathbf{z}_i = \frac{1}{np}\sum^{n}\mathbf{z}_i'P'D^{-1}P\mathbf{z}_i \\ &= \frac{1}{np}\sum^{n}\mathbf{w}_i'D^{-1}\mathbf{w}_i \end{aligned}$$

with $\mathbf{w}_i = P\mathbf{z}_i$, so that $\bar{\mathbf{w}} = \mathbf{0}$ and $S_{\mathbf{w}} = PIP' = I$.Hence

$$1 = \frac{1}{np}\sum^{n}\mathbf{w}_i'D^{-1}\mathbf{w}_i = \frac{\sum^p \lambda_i^{-1}}{p}$$

The result follows now from the well known result that the arithmetic mean is larger than or equal to the geometric mean, with equality iff all the numbers are equal. In fact

$$1 = \frac{\sum^p \lambda_i^{-1}}{p} \geq \left(\prod^p \lambda_i^{-1}\right)^{1/p} \Rightarrow 1 \geq \prod^p \lambda_i^{-1} \Rightarrow 1 \leq \prod^p \lambda_i = \det\left(C\right)$$

with equality iff λ_i^{-1} =constant $(i = 1, 2, ..., p)$, that is, $\lambda_i = 1$, $(i = 1, 2, ..., p)$, and hence $C = I$.

A shorter Proof (using the fact that $(\mathbf{x}, S)$ is the Gaussian MLE)

$$l\left(\boldsymbol{\mu}, \Sigma\right) = c - \frac{n}{2}\log\left(\det\left(\Sigma\right)\right) - \frac{1}{2}\sum^{n}\left(\mathbf{x}_i - \boldsymbol{\mu}\right)'\Sigma^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}\right)$$

We know that

$$l\left(\boldsymbol{\mu}, \Sigma\right) \geq l\left(\bar{\mathbf{x}}, S\right) \tag{2}$$

with equality iff $\boldsymbol{\mu} = \bar{\mathbf{x}}$ and $\Sigma = S$. It is easy to show that, for all Σ,

$$l\left(\boldsymbol{\mu}, \Sigma\right) \leq l\left(\bar{\mathbf{x}}, \Sigma\right) = c - \frac{n}{2}\log\left(\det\left(\Sigma\right)\right) - \frac{1}{2}\sum^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)'\Sigma^{-1}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)$$

Consider the set $\mathcal{C}$ of symmetric, positive definite matrices Σ such that

$$\sum^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)'\Sigma^{-1}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right) = np$$

We know that $S \in \mathcal{C}$. Consider now that problem of minimizing $\log\left(\det\left(\Sigma\right)\right)$ in $\mathcal{C}$. The unique solution to that problem must be $\Sigma = S$. Otherwise we would have found a matrix $\tilde{S}$ such that $l\left(\bar{\mathbf{x}}, S\right) < l\left(\bar{\mathbf{x}}, \tilde{S}\right)$. This would contradict (2).

Minimum Covariance Determinant

The Minimum Covariance Estimator (MCD) was introduced by Rousseeuw (1984). It is defined as follows.

Let $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ be a dataset in R^p. For some integer $0 < h \leq n$ consider all the subsets $\{i_1, i_2, ..., i_h\}$ of $\{1, 2, ..., n\}$ .There are

$$M = \left(\begin{array}{c} n \\ \\ h \end{array} \right)$$

such subsets. Typical values of h are $h = n/2$ or $h = 3n/4$.Set

$$H = \{i_1, i_2, ..., i_h\} \subset \{1, 2, ..., n\} .$$

For each subset H compute the location and scatter matrix

$$\mathbf{t}(H) = \frac{1}{l} \sum_{j \in H} \mathbf{x}_j \quad , \quad S(H) = \frac{1}{h} \sum_{j \in H} (\mathbf{x}_j - \mathbf{t}(H)) (\mathbf{x}_j - \mathbf{t}(H))' .$$

Let

$$H^* = \arg\min_H \det (S(H)) \tag{3}$$

The MCD estimate of multivariate location and scatter is defined as

$$\left(\widehat{\mathbf{t}}, \widehat{\Sigma}\right) = (\mathbf{t}(H^*) \ , S(H^*))$$

The definition of $\left(\widehat{\mathbf{t}}, \widehat{\Sigma}\right)$ is very simple but the computation of H^* in (3) is a highly non-trivial optimization problem.

Next we derive a "concentration step" for the resampling algorithm used to compute MCD.

Lemma 2 *(Rousseeuw and Van Drissen, 1999) Let* $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ *be a dataset in* R^p *and let*

$$H_1 = \{i_1, i_2, ..., i_h\} \subset \{1, 2, ..., n\}$$

of size h *(e.g.* $h = n/2$ *or* $3n/4$*). Let*

$$\mathbf{t}_1 = \frac{1}{h} \sum_{j \in H_1} \mathbf{x}_j \quad , \quad S_1 = \frac{1}{h} \sum_{j \in H_1} (\mathbf{x}_j - \mathbf{t}_1)(\mathbf{x}_j - \mathbf{t}_1)'$$

If $\det(S_1) > 0$ *define the Mahalanobis distances*

$$d_1(i) = \sqrt{(\mathbf{x}_i - \mathbf{t}_1)' S_1^{-1} (\mathbf{x}_i - \mathbf{t})}, \quad i = 1, 2, ..., n.$$

Let now

$$\{k_1, k_2, ..., k_n\}$$

be the permutation of $\{1, 2, ..., n\}$ *such that*

$$d_1(k_1) \leq d_1(k_2) \leq \cdots \leq d_1(k_n)$$

In other words, $d_1(k_l)$ *is the* l^{th} *order statistic of the set* $\{d_1(1), d_1(2), ..., d_1(n)\}$. *Now set*

$$H_2 = \{k_1, k_2, ..., k_h\},$$

$$\mathbf{t}_2 = \frac{1}{h} \sum_{j \in H_2} \mathbf{x}_j \quad , \quad S_2 = \frac{1}{h} \sum_{j \in H_2} (\mathbf{x}_j - \mathbf{t}_2)(\mathbf{x}_j - \mathbf{t}_2)'$$

Then

$$\det(S_2) \quad \leq \quad \det(S_1)$$

with equality if and only if $\mathbf{t}_2 = \mathbf{t}_1$ *and* $S_2 = S_1$.

Proof. (Lemma 2)

Assume, w.l.g., that $\det(S_2) > 0$. Otherwise there is nothing to prove.

Set

$$d_2(i) \quad = \quad \sqrt{(\mathbf{x}_i - \mathbf{t}_2)' S_2^{-1} (\mathbf{x}_i - \mathbf{t}_2)}, \quad i = 1, 2, ..., n.$$

Using the definition of $(\mathbf{t}_2, S_2)$ we write

$$\begin{aligned}
\frac{1}{hp} \sum_{i \in H_2} d_2^2(i) &= \frac{1}{hp} \sum_{i \in H_2} (\mathbf{x}_i - \mathbf{t}_2)' S_2^{-1} (\mathbf{x}_i - \mathbf{t}_2) \\
&= \frac{1}{p} tr \left[\frac{1}{h} \sum_{i \in H_2} (\mathbf{x}_i - \mathbf{t}_2)' S_2^{-1} (\mathbf{x}_i - \mathbf{t}_2) \right] \\
&= \frac{1}{p} tr \left[\frac{1}{h} \sum_{i \in H_2} (\mathbf{x}_i - \mathbf{t}_2)(\mathbf{x}_i - \mathbf{t}_2)' S_2^{-1} \right] \\
&= \frac{1}{p} tr \left[\left(\frac{1}{h} \sum_{i \in H_2} (\mathbf{x}_i - \mathbf{t}_2)(\mathbf{x}_i - \mathbf{t}_2)' \right) S_2^{-1} \right] \\
&= \frac{1}{p} tr \left[S_2 S_2^{-1} \right] = \frac{p}{p} = 1 \qquad (4)
\end{aligned}$$

Now consider

$$
\begin{aligned}
0 &< \lambda = \frac{1}{hp}\sum_{i\in H_2} d_1^2(i) \\
&\leq \frac{1}{hp}\sum_{i\in H_1} d_1^2(i) \quad \text{[by definition of } H_2\text{]} \\
&= \frac{1}{hp}\sum_{i\in H_1} (\mathbf{x}_i-\mathbf{t}_1)' S_1^{-1}(\mathbf{x}_i-\mathbf{t}_1) \\
&= \frac{1}{p}tr\left[\frac{1}{h}\sum_{i\in H_1}(\mathbf{x}_i-\mathbf{t}_1)(\mathbf{x}_i-\mathbf{t}_1)' S_1^{-1}\right] = 1. \qquad (5)
\end{aligned}
$$

Combining (4) and (5) yields

$$
\begin{aligned}
\frac{1}{hp}\sum_{i\in H_2}(\mathbf{x}_i-\mathbf{t}_1)'(\lambda S_1)^{-1}(\mathbf{x}_i-\mathbf{t}_1) &= \frac{1}{\lambda}\frac{1}{hp}\sum_{i\in H_2}(\mathbf{x}_i-\mathbf{t}_1)' S_1^{-1}(\mathbf{x}_i-\mathbf{t}_1) \\
&= \frac{1}{\lambda}\frac{1}{hp}\sum_{i\in H_2} d_1^2(i) \\
&= \frac{\lambda}{\lambda} = 1 \qquad (6)
\end{aligned}
$$

By (4) and (6) we have

$$
\frac{1}{hp}\sum_{i\in H_2}(\mathbf{x}_i-\mathbf{t}_1)'(\lambda S_1)^{-1}(\mathbf{x}_i-\mathbf{t}_1) = \frac{1}{hp}\sum_{i\in H_2}(\mathbf{x}_i-\mathbf{t}_2)' S_2^{-1}(\mathbf{x}_i-\mathbf{t}_2) = 1
$$

and by Lemma 1 we have $\det(S_2) \leq \det(\lambda S_1)$. Moreover,

$$
\det(S_2) \leq \lambda^p \det(S_1) \leq \det(S_1), \quad \text{because } 0 < \lambda \leq 1. \qquad (7)
$$

To have equality in (7), we must have $\lambda = 1$ (otherwise the inequality would be strict). By the uniqueness part in Lemma 1, if $\det(S_2) = \det(\lambda S_1)$ we should have

$$
(\mathbf{t}_2, S_2) = (\mathbf{t}_1, \lambda S_1) \qquad (8)
$$

Hence

$$(\mathbf{t}_2, S_2) = (\mathbf{t}_1, S_1)$$

■

Computing Algorithm for the MCD Estimate

Concentration step. Consider an arbitrary performance criterium denoted $L(\boldsymbol{\theta})$, for example maximum likelihood, MCD, least squares, etc. A *concentration step* is a procedure performed to move from a candidate solution $\boldsymbol{\theta}_0$ to another candidate solution $\boldsymbol{\theta}_1$ that bears a better value for the performance criterium. That is $L(\boldsymbol{\theta}_1)$ is better than $(\boldsymbol{\theta}_0)$.

Lemma 2 justifies the following concentration step regarding the calculation of the MCD estimate.

Given a multivariate sample $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$, a number h (e.g. $h = n/2$) and a candidate solution $(\widehat{\boldsymbol{\mu}}_0, S_0)$, an MCD concentration step is given by the following procedure:

a. Compute the Mahalanobis distances

$$d_0(i) \quad = \quad \sqrt{(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_0)' S_0^{-1} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_0)}, \quad i = 1, 2, ..., n$$

b. Find the subset $H = \{\mathbf{x}_{k_1}, \mathbf{x}_{k_2}, ..., \mathbf{x}_{k_h}\}$ of $\{\mathbf{x}_{k_1}, \mathbf{x}_{k_2}, ..., \mathbf{x}_{k_h}\}$ such that the corresponding h distances $d_0(i)$ are minimized.

c. Compute the sample mean and sample covariance matrix $(\widehat{\boldsymbol{\mu}}_1, S_1)$ for the set H.

The concentration step described above - leading from $(\widehat{\boldsymbol{\mu}}_0, S_0)$ to $(\widehat{\boldsymbol{\mu}}_1, S_1)$ - will be denoted

$$(\widehat{\boldsymbol{\mu}}_1, S_1) \quad = \quad C(\widehat{\boldsymbol{\mu}}_0, S_0).$$

The letter C stands for "concentration step".

Then, by Lemma 2 we have that $\det(S_1) \leq \det(S_0)$.

The Computing Algorithm

Input:

1. the dataset $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$
2. $m = \#$ of subsamples
3. $m_0 = \#$ of top-listed candidates (m_0 is normally much smaller than m)
4. $\delta =$ iterations stopping rule

Output: the MCD estimates $\left(\widehat{\boldsymbol{\mu}}, \widehat{\Sigma}\right)$.

The Steps

1. For $i = 1, 2, ..., m$ extract a random elementary set ($p+1$ randomly chosen sample points), compute the corresponding sample means and covariances $\left(\widehat{\boldsymbol{\mu}}_i^0, \widehat{\Sigma}_i^0\right)$
2. For $i = 1, 2, ..., m$ compute
$$\left(\widehat{\boldsymbol{\mu}}_i^1, \widehat{\Sigma}_i^1\right) = C\left(\widehat{\boldsymbol{\mu}}_i^0, \widehat{\Sigma}_i^0\right)$$
3. Form a subset of m_0 subscripts $\{i_1, i_2, ..., i_{m_0}\}$ for which $\det\left(\widehat{\Sigma}_i^1\right)$ are the smallest.
4. For $j = 1, ..., m_0$ iterate
$$\left(\widehat{\boldsymbol{\mu}}_{i_j}^{k+1}, \widehat{\Sigma}_{i_j}^{k+1}\right) = C\left(\widehat{\boldsymbol{\mu}}_{i_j}^k, \widehat{\Sigma}_{i_j}^k\right)$$
to convergence. Keep the case with smallest $\det\left(\widehat{\Sigma}_{i_j}^\infty\right)$ and report the corresponding $\left(\widehat{\boldsymbol{\mu}}_{i_j}^\infty, \widehat{\Sigma}_{i_j}^\infty\right)$ as the desired solution $\left(\widehat{\boldsymbol{\mu}}, \widehat{\Sigma}\right)$.

S-estimates of Multivariate Location and Scatter

S-estimates for multivariate location and scatter were first defined by Davies (1987).

Given a multivariate sample $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$, let $\rho : [0, \infty) \to [0, \infty)$ be a non-decreasing function such that $\rho(\infty) = \lim_{t\to\infty} \rho(t) = 1$. For each $\mathbf{m} \in R^p$ and each symmetric, positive definite $p \times p$ matrix Σ, let $S(\mathbf{m},\Sigma)$ be defined as the solution in s to the equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{d(\mathbf{x}_i, \mathbf{m},\Sigma)}{sc_p}\right) = \frac{1}{2}$$

where

$$d(\mathbf{x}_i, \mathbf{m},\Sigma) = (\mathbf{x}_i - \mathbf{m})' \Sigma^{-1} (\mathbf{x}_i - \mathbf{m}),$$

is the square Mahalanobis distance from $\mathbf{x}_i$ to $\mathbf{m}$, with covariance matrix Σ. The constant c_p is defined by the equation

$$E\left(\rho\left(\frac{X}{c_p}\right)\right) = \frac{1}{2}$$

with $X \sim \chi_p^2$.

The S-estimate of multivariate location and scatter is now defined in two steps. First define the multivariate location and scatter shape pair $\left(\widehat{\boldsymbol{\mu}}, \widetilde{\Sigma}\right)$ as follows

$$\left(\widehat{\boldsymbol{\mu}}, \widetilde{\Sigma}\right) = \arg \min_{\mathbf{m},\det(\Sigma)=1} S(\mathbf{m},\Sigma).$$

Second, let

$$\widehat{s} = S\left(\widehat{\boldsymbol{\mu}}, \widetilde{\Sigma}\right)$$

Finally set

$$\widehat{\Sigma} = \widehat{s}\widetilde{\Sigma}.$$

References

Davies, P. (1987), "Asymptotic Behaviour of S-Estimates of Multivariate Location Parameters and Dispersion Matrices", Ann. Statist., 15, 1269-1292.

Grubel, R. (1988). "A minimal characterization of the covariance matrix". Metrika, 35, 49-52.

Rousseeuw, P.J. and Van Drissen, K. (1999). "A fast algorithm for the minimum covariance determinant estimator", Technometrics, 41, 212-223

Rousseeuw, P.J. (1984). "Least median of squares regression". journal of the American Statistical Association, 79, 871-880.