Covariance Matrix Estimation for Sparse Data

Jason Bono

November 16, 2022

1 Definitions and Framework

Say we have data sample of k columns and V rows, and with null entries non-trivially distributed throughout. Call the set of null values in the ith column $\emptyset_i$. Call $\mu = \mu_1, \mu_2, ..., \mu_k$ the k dimensional vector of arithmetic means of the latent distribution. The ith element of μ is estimated from the sample by

$$\mu_i = \frac{1}{N_i} \sum_{v \notin \emptyset_i}^{N_i} y_i^v \tag{1}$$

where y_i^v is the vth non-null value (row) of the ith field (column), and N_i is the number of non null values in the ith field. Call ΔA the standard error matrix for any matrix A. The elements of standard error on μ is approximated by

$$((\Delta\mu)_i)^2 = (\Delta\mu_i)^2 \approx (\frac{1}{N_i})^2 \sum_{v \notin \emptyset_i}^{N_i} (y_i^v - \mu_i)^2 \equiv (\frac{\sigma_i}{N_i})^2 \tag{2}$$

or

$$\Delta\mu_i \approx \frac{\sigma_i}{\sqrt{N_i}} \tag{3}$$

Call the latent distribution's covariance matrix R. The ijth element of R can be estimated by

$$\boxed{R_{ij} \approx \frac{1}{n_{ij}} \sum_{v \notin (\emptyset_i \cup \emptyset_j)}^{n_{ij}} (y_i^v - \mu_i)(y_j^v - \mu_j)} \tag{4}$$

where n_{ij} is the number of rows with non-null values in *both* columns i and j. In other words, n_{ij} is the cardinality of the intersection of the ith and jth sets of non null values. Note the sum to n_{ij} only includes rows in the aforementioned intersection.

Note that $n_{ii} = N_i$, so the diagonal terms can be written as

$$\boxed{R_{ii} = \frac{1}{N_i} \sum_{v \notin \emptyset_i}^{N_i} (y_i^v - \mu_i)^2 = \sigma_i^2} \tag{5}$$

The standard error on R is approximated as

$$\Delta R_{ij} \approx \frac{\sigma_i \sigma_j}{\sqrt{n_{ij}}} \tag{6}$$

which, for the diagonal terms, can be written as

$$\Delta R_{ii} = \frac{\sigma_i^2}{\sqrt{n_i}} = \frac{R_{ii}}{\sqrt{n_i}} \tag{7}$$

2 Regularization

The fact that null values are distributed differently from column to column means that the off diagonals R_{ij} are calculated from a subset of data that the diagonals R_{ii} are calculated from. This further implies the possibility of non positive semi definite estimates of R. That is, the best approximations of the individual elements of R_{ij} may not yield a positive semi definite collective form.

In particular since R is symmetric it is guaranteed to be positive semi definite if

$$\psi(R_i) = R_{ii} - \sum_{i \neq j}^{k} |R_{ij}| \geq 0 \tag{8}$$

for all k values of i.

The regularization procedure proposed here is to find the minimal perturbation to R, yielding R', such that, for all negative $\psi(R_i)$, we get

$$\psi(R'_i) = 0 \tag{9}$$

This is achieved for

$$\boxed{\begin{aligned} R'_{ii} &= R_{ii} + a_{ii} \\ R'_{ij} &= R_{ij} - \operatorname{sign}(R_{ij}) a_{ij} \end{aligned}} \tag{10}$$

with any a_i such that

$$\sum_j a_{ij} = -\psi(R_i) \tag{11}$$

An advantageous and unique solution is found by imposing the constraint that the perturbation a_i is distributed across the vector R_i in proportion to the natural variation on the elements R_{ij} . That is,

$$\begin{aligned} a_{ij} &= -\frac{\psi(R_i)}{\sum_j^k (\Delta R_{ij})^2} (\Delta R_{ij})^2; \qquad && \text{if } \psi(R_i) < 0 \\ a_{ij} &= 0; \qquad && \text{otherwise} \end{aligned} \tag{12}$$

where the non trivial case can be written as

$$\boxed{a_{ij} = \frac{-\psi(R_i)}{\sum_j^k (\sigma_i^2 \sigma_j^2 / n_{ij})} (\sigma_i^2 \sigma_j^2 / n_{ij}) = \frac{-\psi(R_i)}{\sum_j^k (\sigma_j^2 / n_{ij})} (\sigma_j^2 / n_{ij})} \tag{13}$$

The advantage to this approach is it guarantees that R' is positive semi definite, it only perturbs R when its estimate is non positive semi definite, in the case of a perturbation the degree of perturbation is minimized, and the parameters that have the smallest standard error get the smallest perturbations.

3 Comparison to R_4^0

In [1], four estimates of the covariance matrix are discussed for initialization in two iterative methods. Here we compare properties of our R with R_4^0 from [1]. The elements of R_4^0 are found by

$$(R_4^0)_{ij} = \frac{1}{\sqrt{N_i N_j}} \sum_{v \notin (\emptyset_i \cup \emptyset_j)}^{n_{ij}} (y_i^v - \mu_i)(y_j^v - \mu_j) \tag{14}$$

Comparison with the unregularized covariance matrix R is straightforward:

$$\frac{R_{ij}}{(R_4^0)_{ij}} = \frac{\sqrt{N_i N_j}}{n_{ij}} \tag{15}$$

For the diagonals

$$\frac{\sqrt{N_i N_i}}{n_{ii}} = 1 \tag{16}$$

For the off diagonals, assume the rows in columns i and j have a probability P of being non null. This is the fraction of rows filled out in i and j, which for simplicity we assume to be identical. In this case,

$$n_{ij} = kP^2 \tag{17}$$

and

$$N_i = N_j = kP \tag{18}$$

so that

$$\frac{\sqrt{N_i N_j}}{n_{ij}} = \frac{1}{P} \tag{19}$$

and therefore

$$\begin{aligned} R_{ii} &= (R_4^0)_{ii} \\ R_{ij} &= \frac{(R_4^0)_{ij}}{P} \end{aligned} \tag{20}$$

so the off diagonal terms in R are larger than R_4^0 in by a factor of $\frac{1}{P}$.

Remember that the regularized R' is either equivalent to R or minimally perturbed to ensure it is positive definite, and that the perturbations are distributed in any offending row (or column) in proportion to the squares of the standard errors.

4 Utilizing the Correlation Coefficient

fill out

References

[1] William J. J. Roberts, *Application of a Gaussian, Missing-Data Model to Product Recommendation* IEEE Signal Processing Letters, Vol. 17, No. 5, 2010.