SEPARABLE MODELS FOR TWO-WAY ARRAYS IN R

JAN DE LEEUW AND WEI TAN TSAI

1. Problem

The $n \times m$ data matrix Y is supposed to be a realization of a matrix normal random variable $\underline{Y}$ [Gupta and Nagar, 2000]. Thus the negative log-likelihood is of the form

$$\mathcal{D}(\mathcal{M}, \Sigma, \Omega) = m \log \mathbf{det}(\Sigma) + n \log \mathbf{det}(\Omega) + \mathbf{tr}\, \Sigma^{-1}(Y - \mathcal{M})\Omega^{-1}(Y - \mathcal{M})'.$$

This can also be written, using Kronecker products, as

$$\mathcal{D}(\mathcal{M}, \Sigma, \Omega) = \log \mathbf{det}(\Sigma \otimes \Omega) + (y - \mu)'(\Sigma \otimes \Omega)^{-1}(y - \mu),$$

where $y = \mathbf{vec}(Y)$ and $\mu = \mathbf{vec}(\mathcal{M})$. See, for example, Abadir and Magnus [2005].

1.1. **Parametrization.** In this paper we choose the *triple-product* or *Pothoff-Roy* parametrization $\mathcal{M} = XBZ'$ for the $n \times m$ matrix of expectations [Pothoff and Roy, 1964]. Matrix X of *row-regressors* is $n \times s$, matrix Z of *column-regressors* is $m \times t$, and matrix B of *regression coefficients* is $s \times t$. In general there may be additional constraints on each of the three matrices in the product.

The dispersions are parametrized as

$$\Sigma \otimes \Omega = \sigma^2(DV(\theta)D) \otimes (EW(\xi)E)$$

Here D and E are diagonal matrices of *row and column variances*, and $V(\theta)$ and $W(\xi)$ are matrix-valued functions defining *row* and *column covariance structures*. Additional constraints are usualluy

Date: Saturday 10th April, 2010 — 12h 4min — Typeset in LUCIDA BRIGHT.

needed for identification, or to arrive at more interpretable structures.

Thus we discuss minimization of the loss function

$$\mathcal{D}(Y,X,B,Z,D,E,\theta,\xi,\sigma^2) = nm\log\sigma^2+m\sum_{i=1}^{n}\log d_i^2+n\sum_{j=1}^{m}\log e_j^2+$$
$$+ m\log\mathbf{det}(V(\theta)) + n\log\mathbf{det}(W(\xi))+$$
$$+\frac{1}{\sigma^2}\mathbf{tr}\ D^{-1}V^{-1}(\theta)D^{-1}(Y-XBZ')E^{-1}W^{-1}(\xi)E^{-1}(Y-XBZ')'$$

over all its nine sets of parameters, with various constraints on each set. The problem is called *separable*, because both in the expectations and in the dispersions there are row and column parameters. No parameters, except perhaps B and σ^2 refer to both rows and columns.

Although our original motivation is the matrix normal distribution, and the negative log likelihood function, we can alternatively just think of our loss function as a way to simultaneously fit simultaneously fit parametric structures to both the expectations and the residuals. The negatve log-likelihood function measures the distance between the data and the covariance matrix of the residuals and the nonlinear manifolds that describe their expectations [de Leeuw and Kreft, 1986]. That distance is then minimized, i.e. data and residuals are projected on these manifolds.

2. Constraints

2.1. **Expectations.** The matrices X, B and Z can be partially fixed and partially free, which means minimization is only over a subset of their elements. The free elements can be restricted to be non-negative.

If X and Z are free and B is fixed to the identity we fit variations of *principal component analysis* (and *non-negative PCA* if we require the elements of X and Z to be non-negative). If X and Z are both

fixed and only B is free we fit *growth curve models*, if X is fixed and B and Z are free this becomes *redundancy analysis* or *reduced rank regression analysis*.

2.2. **Dispersions.** D and E are (positive) diagonal matrices. They can be either free, and we minimize over them, or they can be fixed to given matrices. They are fixed, for example, in *simple, multiple, canonical, and joint correspondence analysis* [Greenacre and Blasius, 2006], in which we choose them equal to the marginals of the contingency table or Burt table.

Matrix V is a (positive definite) matrix-valued function of $p \geq 0$ parameters θ, and matrix W is a (positive definite) matrix-valued function of $q \geq 0$ parameters ξ. We define the matrices of partial derivatives $G_s(\theta) \triangleq \mathcal{D}_s V(\theta)$ and $H_s(\xi) \triangleq \mathcal{D}_s W(\xi)$. One options is that V and/or W are equal to fixed matrices, in which case the partial derivatives are obviously equal to zero.

One example of the parametric specification of V is the *linear covariance structure*

$$V(\theta) = I + \sum_{s=1}^{p} \theta_s G_s,$$

with the G_s known matrices. Obviously in this case $G_s(\theta) = G_s$. An important special case is $V(\theta) \equiv I$.

Another example is the *common factor structure*

$$V(\theta) = I + A(\theta)A(\theta)',$$

where $A(\theta) = \sum_{s=1}^{p} \theta_s U_s$, with the U_s known matrices. Here $G_s(\theta) = A(\theta)U_s' + U_s A(\theta)'$.

There are also one-parameter structures, such as the *exponential distance structure* with

$$v_{ik}(\theta) = \begin{cases} 1 \text{ if } i = k, \\ \exp(-\theta\delta_{ik}) \text{ otherwise.} \end{cases}$$

where Δ is a given distance-like (symmetric, positive, and hollow) matrix. A special case is the *Kac-Murdoch-Szego structure*, which has $\delta_{ik} = |i - k|$, or the *Olkin-Press structure*, which has $\delta_{ik} = \min(|i - k|, n - |i - k|)$. These structures are scaled such that $v_{ik}(0) = 1$ for all i, k and $\lim_{\theta\to\infty} V(\theta) = I$.

2.3. **Data.** The data Y can have missing components, i.e. some elements may be unknown. We have implemented two ways to deal with such *missing data*. The empty cells can be additional parameters over which the loss function is minimized, or they can be imputed by the multinormal EM algorithm. In the context of point estimation the two methods are compared in Little and Rubin [1983].

3. Algorithm

The basic algorithm is *block relation* [Oberhofer and Kmenta, 1974; De Leeuw, 1994]. We cycle through the nine blocks of parameters, and in each of the nine sub-steps that define a cycle we keep eight blocks of parameters fixed at their current values and minimize over the ninth blocks.

Of course in some cases blocks of parameters are fixed, and we can skip the corresponding substep. In other cases minimizing over a parameter block is an iterative process which does not converge in a finite number of iterations, and we must truncate it before it is properly finished.

For the minimization in the substeps we use the R function `nlminb`, which also allows for bound constraints on the parameters in the block. In the special case of one-parameter models for $V(\theta)$ and $W(\xi)$ we can also use the `optimize` function.

Define

$$\hat{Y}(X,B,Z) \stackrel{\Delta}{=} XBZ',$$
$$R(Y,X,B,Z) \stackrel{\Delta}{=} Y - \hat{Y}(X,B,Z),$$
$$\overline{V}(\theta,D) \stackrel{\Delta}{=} DV(\theta)D,$$
$$\overline{W}(\xi,E) \stackrel{\Delta}{=} EW(\theta)E.$$

4. Step 1: Minimization over σ^2

The loss function is minimized at

$$\hat{\sigma}^2 = \frac{1}{nm}\mathbf{tr}\,\overline{W}^{-1}R'\overline{V}^{-1}R.$$

5. Step 2: Minimization over D and E

If D is free, then we must minimize

$$f(d) = m\sum_{i=1}^{n}\log d_i^2 + \mathbf{tr}\,D^{-1}V^{-1}D^{-1}S$$

where $S = R\overline{W}^{-1}R'$. Suppose $T = S \star V^{-1}$ is the elementwise (Hadamard) product of S and V^{-1}. Then

$$f(d) = 2m\sum_{i=1}^{n}\log d_i + \sum_{i=1}^{n}\sum_{k=1}^{n}\frac{t_{ik}}{d_i d_k},$$

with derivative

$$\mathcal{D}_i f(d) = 2\frac{m}{d_i} - 2\frac{1}{d_i^2}\sum_{k=1}^{n}\frac{t_{ik}}{d_k}.$$

Obviously the equations for minimizing over E are similar.

6. Step 3: Minimization over V and W

The part of the loss function that depends on θ is

$$f(\theta) = m\log\mathbf{det}(V(\theta)) + \mathbf{tr}\,V^{-1}(\theta)S,$$

where

$$S = D^{-1}R\overline{W}^{-1}R'D^{-1}$$

The partial derivatives are

$$\mathcal{D}_s f(\theta) = m\mathbf{tr}\, V^{-1}(\theta)G_s(\theta) - \mathbf{tr}\, V^{-1}(\theta)G_s(\theta)V^{-1}(\theta)S.$$

7. Step 4: Minimization over X, B, Z

We have to minimize

$$S(X) = \mathbf{tr}\, V^{-1}(Y - XBZ')W^{-1}(Y - XBZ')'$$

The unconstrained optimum is attained at any solution $\tilde{X}$ of

$$X[BZ'W^{-1}ZB'] = YW^{-1}ZB'.$$

Define $C = BZ'W^{-1}ZB'$. Partition the residual sum of squares by using

$$S(X) = S(\tilde{X}) + \mathbf{tr}\, (X - \tilde{X})'V^{-1}(X - \tilde{X})C.$$

Thus if there are constraints on X we must minimize the second term

$$\mathcal{P}(X) = \mathbf{tr}\, (X - \tilde{X})'V^{-1}(X - \tilde{X})C. \tag{1}$$

7.1. **Pattern.**

References

K.M. Abadir and J.R. Magnus. *Matrix Algebra*. Cambridge University Press, Cambridge, GB, 2005.

J. De Leeuw. Block Relaxation Methods in Statistics. In H.H. Bock, W. Lenski, and M.M. Richter, editors, *Information Systems and Data Analysis*, Berlin, 1994. Springer Verlag.

J. de Leeuw and I. G. G. Kreft. Random Coefficient Models for Multilevel Analysis. *Journal of Educational Statistics*, 11:57–85, 1986.

M. Greenacre and J. Blasius, editors. *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall, 2006.

A.K. Gupta and D.K. Nagar. *Matrix Variate Distributions*. Chapman & Hall/CRC, Boca Raton, Florida, 2000.

R.J.A. Little and D.B. Rubin. On Jointly Estimating Parameters and Missing Data by Maximizing the Complete Data Likelihood. *American Statistician*, 37:218–220, 1983.

W. Oberhofer and J. Kmenta. A General Procedure for Obtaining Maximum Likelihood Estimates in Generalized Regression Models. *Econometrica*, 42:579–590, 1974.

R. F. Pothoff and S. N. Roy. A Generalized Multivariate Analysis of Variance Model Useful Especially for Growth Curve Problems. *Biometrika*, 51:313–326, 1964.

Department of Statistics, University of California, Los Angeles, CA 90095-1554

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://gifi.stat.ucla.edu

E-mail address, Wei Tan Tsai: tsai@stat.ucla.edu

URL, Wei Tan Tsai: http://www.stat.ucla.edu/~tsai