Chapter 1. Descriptive Techniques

Chapter 9. Decomposition of Data Matrices by Factors

* Reduction of dimension using a geometrical approach
* Derived with respect to a least-squares criterion
* Low dimensional graphical pictures of the data matrix
  + decomposition of data matrix into “factors” in decreasing order
  + general idea is the core of many multivariate techniques
    - Ch 10: Principal Components (linear combinations)
    - Ch 11: Factor Analysis (fixed number of factors – latent characteristic)

9.1 Geometric Point of View

* Data matrix **X**n x p is a cloud of *p* points in **R***n*.
  + rows: (*n* points in )
  + columns: (*p* points in )
  + when *n* and/or *p* > 2 or 3, no interpretable graphs; therefore aim of factorial methods is to simultaneously approximate column and row spaces with smaller subspaces

9.2 Fitting the *p*-dimensional Point Cloud (row space)

* project the point cloud onto a space of lower dimension
  + subspace of dimension 1 (straight line F1 through the origin with direction unit vector )
    - *xi* on unit vector is represented by projection
    - *best line F1* minimizes
    - that is, maximizes under the constraint.
* Theorem 9.1: *The vector u1 which minimizes F1 is the eigenvector of* ***X****T****X*** *associated with the largest eigenvalue λ1 of* ***X****T****X****.*
  + When the data is centered, then Xc is the centered data matrix, and 1/*n* ***X****T****X*** is the covariance matrix
* Representation of the Cloud on F1
  + coordinates of the *n* individuals *xi* are represented by new factorial variable where *x[i]* are the original variables.
  + ***X****u1* is the *first factorial variable* or *first factor* and *u1* is the *first factorial axis*
* Subspaces of Dimension 2
  + If the *n* individuals are approximated by a plane, this space contains *u1*.
  + The plane is determined by best linear fit *u1* and a unit vector *u2* which maximizes under the constraints and
* Theorem 9.2: *The second factorial axis, u2, is the eigenvector of* ***X****T****X*** *associated with the second largest eigenvalue λ2 of* ***X****T****X****.*
* Subspaces of dimension *q* (*q* ≤ *p*)
  + The arguments above can be extended to the orthonormal eigenvectors of ***X****T****X***.

9.3 Fitting the *n*-dimensional point cloud

* Subspaces of Dimension 1
  + subspace of dimension 1 (straight line G1 through the origin with direction unit vector )
    - *xi* on unit vector is represented by projection
    - *best line F1* minimizes
    - that is, maximizes under the constraint.
* Theorem 9.3: *The vector v1 which minimizes G1 is the eigenvector of* ***XX****T associated with the largest eigenvalue ν1 of* ***XX****T.*
* Representation of the cloud on G1.
  + as above for F1
* Subspaces of dimension *q* (*q* ≤ *n*)
  + as above for F1

9.4 Relations between subspaces

* each eigenvector of **X**T**X** corresponds to an eigenvector of **XX**T associated with the same eigenvalue
* Theorem 9.4 (Duality Relations) *Let r be the rank of* ***X****. For k ≤ r, the eigenvalues λk of* ***X****T****X*** *and* ***XX****T are the same, and the eigenvectors (uk and vk, respectively) are related by*
* Recall
  + coordinates representing the columns of **X** in a *q*-dimensional subspace can be easily calculated by

9.5 Practical computation

* **X** is often centered and scaled (with MLE standard deviation, not sample, i.e. )
* begin with computing eigenvalues (λi) and corresponding eigenvectors (ui) of **X**T**X**.
* plot vs (*n* points)
* plot vs (*p* points)
* evaluate the quality of the factorial representations in a subspace of dimension *q* by , the percentage of the inertia (yTy) explained by the first *q* factors

Chapter 10. Principal Components Analysis

* Used to reduce the dimension of X
* Aim is to find linear combinations which create the largest variances
* Not scale invariant: Standardize measurement units

10.1 Standardized Linear Combination

* such that (that is, δ is a unit vector) { this is an SLC }
* maximize variance of projection by choosing δ which satisfies
* δ is given by the eigenvector corresponding to the largest eigenvalue λ of the covariance matrix.
* Principal Components
  + PC1:
  + PC2:
* For X such that E[X] = μ and Var(X) = Σ = ΓΛΓT
* Theorem 10.1. *For a given . Let be the PC transformation. Then*
  + Var(Y­1) ≥ Var(Y­2) ≥ … ≥ Var(Y­p) ≥ 0
* Theorem 10.2. *There exists no SLC that has a larger variance than λ1 = Var(Y1).*
* Theorem 10.3. *If is an SLC that is not correlated with the first k PCs of X, then the variance of Y is maximized by choosing it to be the (k+1) PC.*

10.2 Principal Components in Practice

* μ is estimated by ; Σ is estimated by SMLE = S \* (n-1)/n
* If S = VDVT is the spectral decomposition of S, then the PCs are obtained by:
* WARNING: PC analysis should be applied to data that have approximately the same scale in each variable.

10.3 Interpretation of the PCs

* The proportion of variance explained is
* A measure of how well the first *q* PCs explain variation is given by
* Example 1:
  + The PC is essentially the difference between x4 and x6.
* Example 2:
  + The PC is essentially the different between x­5­ and the sum of x­4 and x6.
* A scree plot gives a good graphical representation of the ability of the PCs to explain the variation in the data

Chapter 13. Discriminant Analysis

* Used when clusters (classes) are known *a priori* (sometimes by cluster analysis on past data)
* aim is to classify observation(s) into known groups

13.1 Allocation rules for known distributions

* credit scoring for bank loans
* counterfeit bank notes
* “fast” and “slow” customers of a newly introduced product
* attribution of unknown literary or artistic works to know writers/artists
* a *discriminant rule* is a separation of the sample space (usually **R**p) into sets Rj such that the error of misclassification is small
* **Maximum Likelihood Discriminant Rule**
  + let each population , then the *maximum likelihood discriminant rule* is given by allocating *x* to Пj maximizing the likelihood
  + probability of misclassification errors (putting *x* into *i* when it is from population *j*) can be calculated as
  + misclassifications create a cost C(*i|j*) when an observation from Пj is assigned to Rj.
  + if πj is the *a priori* probability that an individual selected at random belongs to Пj, then the *expected cost of misclassification* ECM is given by .
  + Classification rules keep the ECM small, or minimized over a class of rules
* Theorem 13.1. *For two given populations, the rule minimizing the ECM is given by*
* Theorem 13.2. *Suppose . Then*
  + *The ML rule allocates x to Пj, where is the value minimizing the square Mahalanobis distance between x and μi:*
  + *In the case of J = 2, , where and*
* **Bayes Discriminant Rule**
  + the Bayes rule of discrimination allocates *x* to the Пj that gives the largest values of πj*fi*(*x*). It is identical to the ML discriminant rule for πj = *1/J.*
  + a *randomized discriminant rule* allocates *x* to Пj with a certain probability *φj(x)*.
  + comparing discriminant rules: let the probability of allocating *x* to Пj correctly be . Then, a discriminant rule with probabilities *p­ij* is as good as any other discriminant rule with probabilities if for all *i = 1, …, J.* The first rule is better if the strict equality holds for at least one *i*.
  + a discriminant rule is called *admissible* if there is no better discriminant rule.
* Theorem 13.3. *All Bayes discriminant rules (including the ML rule) are admissible.*
* **Probability of misclassification for the ML rule (J = 2)**
  + for (equal variance)
  + If and where is the squared Mahalanobis distance between the populations, then
* **Classification with different covariance matrices**
  + when covariance for both density functions differ, the allocation rule is
  + where
  + These belong to the family of Quadratic Discriminant Analysis

13.2Discrimination rules in practice

* Common Covariance: The ML rule is used when the distribution is known up to parameters (*e.g.* multivariate normal). For *J* groups with *nj* observations in each group, we use   to estimate *μj* and *Sj* to estimate *Σj*, so that the empirical version of the ML rule is to allocate a new observation *x* to *Пj* such that *j* minimizes:
  + , with estimates the common covariance
* Calculate the hyperplane separating populations with
  + and
  + Count misclassifications by
    - note summation over first group
    - summation over second group
  + For J = 3 groups, allocation regions can be calculated using
    - allocate *x* to
* **Estimation of the probabilities of misclassifications**
  + estimate by replacing unknown parameters by estimators
  + ML rule for two normal populations
    - where
  + *re-substitution method* (leads to too optimistic estimators)
    - reclassify original observations according to chosen rule
    - ; the matrix is known also as the *confusion matrix*
    - APER { Apparent Error Rate } is the percentage of observations that are misclassified
  + approach to correct for bias introduced by using observations twice (once to construct rule, once to evaluate) in re-substitution rule above
    - For each population separately:
      * develop classification rule on *n1 -1, n2* observations and classify the holdout based on this rule. Repeat until all population 1 are classified. Count the number of *n1’* misclassifications*.* Repeat for population 2; count the number of *n2’* misclassifications.
      * ;
    - a better estimator of the actual error rate (AER) is
* **Fisher’s Linear Discrimination Function**
  + For **Y** = **Xa**, a linear combination of observations, then the total sum of squares is
    - * is the centering matrix
  + Find the *a*T*x* which maximizes the ratio of the *between-group-sum of squares* to the *within-group-sum of squares*.
    - { within group }
    - { between group }
    - { total SS }
  + Theorem 13.4. *The vector a that maximizes is the eigenvector of W-1B that corresponds to the largest eigenvalue.*
  + Classification rule:
  + For *J = 2* groups
  + closely related to projection pursuit (Chapter 19)

Chapter 14.

Chapter 15. Canonical Correlation Analysis

* Joint study of wo data sets: ask what type of low-dimensional projection helps to find joint structures
* Standard tool for discovery and quantification of associations between two sets of variables (aims to identify possible links between two sets of variables)
* Basic technique: define an index (projected multivariate variable) that maximally correlates with the index of the other variable for each sample separately.
* Aim: maximize the association (correlation) between the low-dimensional projections of the two data sets

15.1 Most Interesting Linear Combination

* CCA developed by Hotelling (1935)
* for random variables and find such that is maximized.
* given
* ; covariance is invariant to scale, so we can solve
* under the constraints
* define , then the SVD is (or UΛVT)
* where K has rank k = rank() = rank(), Λ is the diagonal matrix with non-zero eigenvalues (sqrt) of KKT and KTK, Γ are the standardized eigenvectors of KKT, and Δ are the standardized eigenvectors of KTK.
* The *canonical correlation vectors* are:
* The *canonical correlation variables* are:
* The *canonical correlation coefficients* are:
  + for *i* = 1, …, *k*
* The canonical correlation variables are independent:
* Theorem 15.1 *For any given r, 1 ≤ r ≤ k, the maximum subject to and for i, …, r-1 is given by and is attained when a = ar and b = br.*
  + Thus, the canonical correlation vectors calculated with the eigenvectors associated with the largest eigenvalues maximize the correlation between the canonical variables associated with the largest eigenvectors.
* Theorem 15.2. *Let ηi and ϕi be the ith canonical correlation variables (i = 1, …, k). Then*  where .
* Theorem 15.3. *Let and where* ***U*** *and* ***V*** *are nonsingular matrices. Then the canonical correlations between* ***X\**** *and* ***Y\**** *are the same as those between* ***X*** *and* ***Y****. The canonical vectors of* ***X\**** *and* ***Y\**** *are given by and .*