Data: $Y = \{y_{ij} : i=1\ldots m_1,\ j=1\ldots m_2\}$ $m_1 \geq m_2$

SVD: $Y = UDV^T$, $U \in \mathcal{V}_{m_2,m_1}$, $V \in \mathcal{O}_{m_2}$, $D = \text{diag}(d_1 \ldots d_{m_2})$

$y_{ij} = u_i^T D v_j$

Factor model: $y_{ij} = a_i^T b_j + \epsilon_{ij}$ → truncated SVD gives LS est.

Comparison to ANOVA:

Consider a 2-factor layout, one obs per factor combination

y_{ij} = obs under factor $1 = i$, factor $2 = j$

Ex:

variety \ fert	1	2	3
1	y_{11}	$\cdot$	–
2	$\cdot$	y_{22}	y_{23}
3	$\cdot$		$\cdots$
4			

y_{ij} = crop yield for variety i, fert. j

Questions: what is the best fertilizer?
what is the best variety?

Model: $y_{ij} = \mu + a_i + b_j + \epsilon_{ij}$ #

i_1 is better than i_2 if $a_{i_1} > a_{i_2}$
j_1 " j_2 $b_{j_1} > b_{j_2}$

is a linear model OLS/MLEs given by

$\hat{\mu} = \bar{y}_{\cdot\cdot}$

$\hat{a}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$

$\hat{b}_j = \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}$

$$y_{ij} = \mu + a_i + b_j + \epsilon_{ij}$$
$$= \underbrace{\bar{y}_{\cdot\cdot}}_{\hat{\mu}} + \underbrace{(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})}_{\hat{a}_i} + \underbrace{(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})}_{\hat{b}_j} + \underbrace{(y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})}_{\hat{\epsilon}_{ij}}$$

Fitted Values: $\hat{y}_{ij} = \hat{\mu} + \hat{a}_i + \hat{b}_j$, $\hat{Y} = \hat{\mu}11^T + \hat{a}1^T + 1\hat{b}^T$

What is the rank of $\hat{Y}$? # Go through example in R*

1, 2, 3, $m_1 \vee m_2$?

$\hat{Y} = \hat{u}\,\mathbb{1}\mathbb{1}^T + \hat{a}\mathbb{1}^T + \mathbb{1}\hat{b}^T \rightarrow$ sum of 3 rank 1 matrices, rank ≤ 3

$= (\hat{u}\underset{\sim}{\mathbb{1}} + \underset{\sim}{\hat{a}})\mathbb{1}^T + \mathbb{1}\hat{b}^T \rightarrow$ sum of 2 rank 1 mt., rank $\leq$

can show rank is (generally) $\underline{\underline{2}}$

Questions

$\min_{\underset{\sim}{a},\underset{\sim}{b},\mu} \| Y - (\mu + a\mathbb{1}^T + \mathbb{1}b^T) \|^2 \ \substack{>\\<} \ \min_{M:\mathrm{rank}(M)=2} \| Y - M \|^2$

(2) which model has more parameters (additive eff vs SVD mult model)?

$\left(\dim(\mathcal{V}_{kn}) = nk - \frac{1}{2}k(k+1) \right)$

(3) Discuss merits/problems w/ each model (add vs. mult)

$k=2$

$$\overset{u}{(2m-3)} + \overset{v}{(2n-3)} + \overset{d}{2} = \underline{\underline{2(m+n)}}-4 \quad \text{vs} \quad (m-1)+(n-1)+1$$

$$(2(m_1+m_2)-4) \qquad \underline{\underline{m+n-1}}$$

$$m_1+m_2-1$$

$2(m+n)-4 > m+n-1$

$\underline{\underline{(m+n)}} > 3$ $\qquad m+n=3$ means only 1 factor

$\rightarrow$ additive model always simpler

Evaluating Additivity

Classical approach : Y_{ijk} = kth rep under (i,j)

H : $Y_{ijk} = \mu + a_i + b_j + \epsilon_{ijk}$

K : $Y_{ijk} = \mu + a_i + b_j + c_{ij} + \epsilon_{ijk}$ $\quad c_{ij}$ is the "interaction" between the two factors

if $k > 1$, H vs K can be evaluated w/ standard F-test.

obs : mn
pars in add : $m+n-1$ ⟷ can potentially check for something in between.
pars in full : mn

What if $K=1$?

(Tukey 1949) $H: \; Y_{ij} = \mu + a_i + b_j + \epsilon_{ij}$
$K: \; Y_{ij} = \mu + a_i + b_j + \lambda a_i b_j + \epsilon_{ij}$

Test: $SS1 = \sum (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$
$SS2 = \sum (\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2$
$SSI = \sum_{ij} Y_{ij} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})$ → under H, $\frac{(SSI)^2}{SS1\,SS2} \sim F_{1, mn-(m+n)}$

large if Y_{ij} "correlated" with $\hat{a}_i \hat{b}_j$.

Problem: what if departure from additivity is not of this form

* Open question: (?) test $H: \; Y_{ij} = \mu + a_i + b_j + \epsilon_{ij}$ (additive)
$K: \; Y_{ij} = \mu + a_i + b_j + d u_i v_j + \epsilon_{ij}$ (rank 1 interaction)

or $H: \; Y_{ij} = \mu + \epsilon_{ij}$
$K: \; Y_{ij} = \mu + d u_i v_j + \epsilon_{ij}$ { LRT, invariant test, (max invariant is available)

-1 = z

zλ+1

Other statistical issues
missing data → do ALS demo.
Non normality / different norms / trans. data { $Y \approx UDV'$ or $Y \approx g(UDV') \Leftrightarrow g^{-1} Y \approx UDV'$ ← consider network data social networks etc.

Box Cox (1964): Power transforms for positive data $y_1 \dots y_n$

$Z_i = \beta^T X_i + \epsilon_i$

$Y_i = (\lambda Z_i + 1)^{1/\lambda}$
$\Leftrightarrow Z_i = \frac{Y_i^{\lambda} - 1}{\lambda}$ (use log/exp if $\lambda = 1$)

[illegible] $p(Y \mid \lambda, \beta, X)$

Open question:

$Z = \{Z_{ij}\} = UDV^T + E$

$Y_{ij} = g(Z_{ij})$

how to estimate g, U, D, V^T simultaneously