k-Means and Grouped FE

Chris Conlon

May 2, 2021

Applied Econometrics II

Reading

- Chapters 14 of *Elements of Statistical Learning*
- Bonhomme, Lamadon and Manresa
  - A Distributional Framework for Matched Employer Employee Data (Econometrica 2020)
  - Discretizing Unobserved Heterogeneity (2021)

k-means Clustering

Idea: I have a matrix of data $\mathbf{X}$ ($N \times M$).

- We are going to take the rows of X and assign them to K groups.
  - Groups are mutually exclusive and exhaustive.
- Goal: minimize residual variance:

$$\min_{\mu} \min_{k(i)} \sum_{i=1}^{N} \sum_{k=1}^{K} ||\mathbf{x_i} - \mu_{k(i)}||^2$$

- Choose both
  - Assignment of each row to a group $k(i)$
  - Mean of each group μ_k.

k-means Clustering

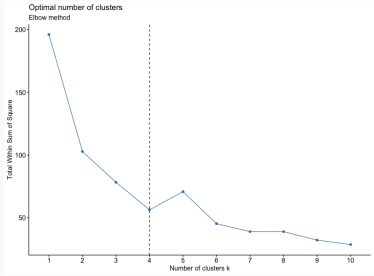
How does it work?

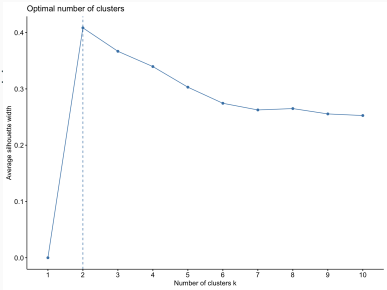
- Mean-Variance tradeoff
  - More clusters: less bias, higher variance
  - Fewer clusters: more bias, less variance.
- Naive idea: alternate between minimum distance to assign groups and recompute mean of group.
- Honestly this is a difficult problem for the computer to solve (NP-hard).
- You don't want to implement it on your own.
- Use a canned routine.
- Choice of distance metric matters: Euclidean L_2, Mahalanobis (Covariance), Manhattan/Taxicab L_1.

How to choose K?

- Tune with OOS MSE to choose K via cross-validation
- Most people use "elbow method" on the right.
- See `https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-cluste`

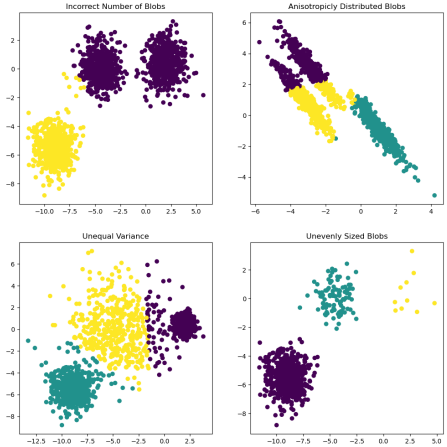
```
library("factoextra")
my_data <- scale(USArrests)
fviz_nbclust(my_data, kmeans, method = "gap_stat")
fviz_nbclust(my_data, kmeans)
```





Not as easy as it looks (figure from `scikit-learn`)

- Assumes that variance is essentially spherical so standardize inputs first!
- Rotations of space (e.g. PCA) will change the answer!
- Scaling of the space (e.g. taking $\log(X)$) will change the answer!
- Choosing wrong number of clusters K is a bad idea.



Why Bother?

- May want to categorize data into distinct groups by similar characteristics
  - High growth firms; "Cash-Cows", etc.
  - Healthy v. Unhealthy, Kids v. Adult Cereals, etc.
  - But groups may not be interpretable..
- Predicting with the group mean has very low variance (probably too low).

Grouped Fixed Effects: Bonhomme, Lamadon, Manresa (2021)

Wages: W_{it} and Y_{it}: Labor Force Participation $\{0, 1\}$.

$$
\begin{aligned}
Y_{it} &= \mathbf{1}\left\{u\left(\alpha_{i0}\right) \geq c\left(Y_{i,t-1}; \theta_0\right) + U_{it}\right\} \\
W_{it}^{*} &= \alpha_{i0} + V_{it} \\
W_{it} &= Y_{it} \cdot W_{it}^{*}
\end{aligned}
$$

- conventional FE estimator would treat α_{i0} and $u(\alpha_{i0})$ as unrelated parameters
- so the FE estimate of θ would be solely based on the binary participation decisions.
- Instead assign a common α_{i0} to individuals with similar W_{it}, Y_{it} vectors.

Grouped Fixed Effects: What's the point?

- Large class of models looking at matched employer-worker data.
- Do high FE workers match to high or low FE firms?
- Goal is often to disentangle firm FE from worker FE.
- But workers don't change firms very often and often absorbed into firm FE.
  - only switchers allow identification
- See:
  - Abowd, Kramarz and Margolis (1999): estimate full FE
  - Card, Heining and Kline (2013): bin workers by quartile of wages.
  - Bonhomme, Lamadon, Manresa (2019): grouped FE.

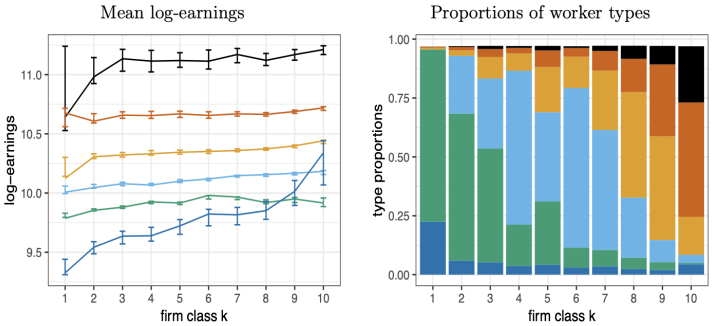
Matched Firm/Worker Example

$$Y_{it} = a_t\left(k_{it}\right) + b_t\left(k_{it}\right)\alpha_i + X_{it}'c_t + \varepsilon_{it}$$

- Worker FE α_i drawn from distribution that depends on firm type k_{it}
- Wages Y_{it} drawn from distribution that depends on α_i and X_{it} as well as firm type k_{it}
- Using k-means:
  - Assign firms to clusters $K = 10$ based on CDF of log earnings.
  - Assign workers to $L = 6$ discrete types.

Grouped FE results

Figure 2: Main parameter estimates of the static model



Notes: Estimates of the static model, on 2002-2004. In the left graph we plot estimates of means of log-earnings, by worker type and firm class. We order the $K = 10$ firm classes (on the x-axis) by mean log-earnings. On the y-axis we report estimates of mean log-earnings for the $L = 6$ worker types. In the right graph we show estimates of the proportions of worker types in each firm class. In the left graph, the brackets indicate pointwise parametric bootstrap 2.5%–97.5% quantile bands (computed using 200 replications).

Grouped FE results

Table 2: Variance decomposition and reallocation exercise in the static model

Variance decomposition ($\times 100$)

$\frac{Var(\alpha)}{Var(y)}$	$\frac{Var(\psi)}{Var(y)}$	$\frac{2Cov(\alpha,\psi)}{Var(y)}$	$\frac{Var(\varepsilon)}{Var(y)}$	$Corr(\alpha,\psi)$
60.03	2.56	12.17	25.24	49.13
(0.85)	(0.16)	(0.39)	(0.59)	(0.86)

Reallocation exercise ($\times 100$)

Mean	Median	10%-quantile	90%-quantile	Variance
0.50	0.58	2.60	-1.24	-1.12
(0.10)	(0.11)	(0.19)	(0.31)	(0.11)

Notes: Estimates of the static model, on 2002-2004. In the top panel, α denotes the worker effect, and ψ denotes the firm effect, in the linear regression $Y = \alpha + \psi + \varepsilon$. In the bottom panel we report differences in means, quantiles, and variances of log-earnings between two samples: a counterfactual sample where workers are randomly reallocated to firms, and the original sample. The results are obtained using $1,000,000$ simulations, and we report parametric bootstrap standard errors in parentheses (computed using 200 replications).

Thanks!