Calibrating on Principal Components in the presence of Multiple Auxiliary Variables for Nonresponse Adjustment

Bernardo Rota

&

Thomas Laitila

Introduction-Motivation

- Auxiliary information plays a prominent role in successful estimation. Rizzo, Kalton and Brick (1996) note that, providing it is carefully chosen, the particular adjustment scheme used at the estimation stage is not that important.
- Increase of storage capacities and automatic data collection process may lead to data analysis of large set of auxiliary variables.
- The use of large sets of auxiliary variables may incur in constructing estimators with worse performances (inefficient than HT, Cardot, Goga and Shehzad, 2014)
- The Computational effort increases with the number of variables
- Common practice of usage of large sets of auxiliary variables is based on selection criteria of subsets deemed important auxiliary variables
- In multipurpose estimations it is less likely that specific variable become important to all study variables or at least all important study variables

Principal Components as an alternative to selection criteria

- Gives an optimal linear combination of all candidate auxiliary variables
- Produce a joint effect of the candidate variables on the variables of interest.
- Reduces the computational effort due to large number of auxiliary variables

Summary on Principal Components

- Let $\mathbf{X}_{(N \times P)}$ be our auxiliary data matrix, and suppose that

$\mathbf{X}_{j(N \times 1)}, \{j = 1, \ldots, P\}$ is rescaled to zero mean and unit variance, then

$\mathbf{X}^t\mathbf{X}$ is the covariance matrix of $\mathbf{X}$. If $(\lambda_j, \mathbf{b}_j; j = 1, \ldots, P)$ are pairs

eigenvalue-eigenvector of $\mathbf{X}^t\mathbf{X}$, then, the j^{th} principal components

is given by $Z_j = \mathbf{b}_j^t\mathbf{X}$ all Z_j's are uncorrelated.

- When $\mathbf{X} = \mathbf{X}_{(n \times P)}$, then, $\operatorname{cov}(\mathbf{X}) = \mathbf{X}^t D \mathbf{X}$

  where $D = diag\left\{d_1, \ldots, d_n\right\},\ d_k = 1/\pi_k$.

  $\left(\hat{\lambda}_j, \hat{\mathbf{b}}_j; j = 1, \ldots, P\right)$ are eigenvalue-eigenvector pairs of $\mathbf{X}^t D \mathbf{X}$ this

  leads to estimated principal components $\hat{Z}_j = \hat{\mathbf{b}}_j^t \mathbf{X}$

Estimators under consideration

- Linear Calibration estimator (Särndal & Lundström, 2005)

$$\hat{t}_{Lc} = \sum_r w_k^{pc} y_k$$

$$w_k^{pc} = d_k + d_k \boldsymbol{\delta}_{(pc)}^t \mathbf{Z}_k \quad \text{and} \quad \boldsymbol{\delta}_{(pc)} = \left(\mathbf{Z}^t \mathbf{D} \mathbf{Z}\right)^{-1} \left(\mathbf{T}_z - \mathbf{Z}^t \mathbf{d}\right)$$

$$\mathbf{D} = diag\left\{d_1, d_2, \ldots, d_k, \ldots, d_m\right\}$$

- Propensity Score Calibration (Chang and Kott, 2008)

$$\hat{t}_{psCal} = \sum_r d_k \phi(\mathbf{x}_k^t \hat{\boldsymbol{\delta}}_{(pc)}) y_k$$

$where:$

$\hat{\boldsymbol{\delta}}$ is solution to $\sum_r d_k \phi(\mathbf{x}_k^t \hat{\boldsymbol{\delta}}_{(pc)}) \mathbf{z}_k = \mathbf{T}_z$

Components Retention

- Selecting R components among P is a theme that has widely been considered, e.g. Jolliffe (1972, 1973, 1982), Cadima and Jolliffe (1995), Jolliffe, Trendalov, and Uddin (2003), and McCabe(1984).

- A Canonical correlation between $\mathbf{H}$ and $\tilde{\mathbf{Z}}$

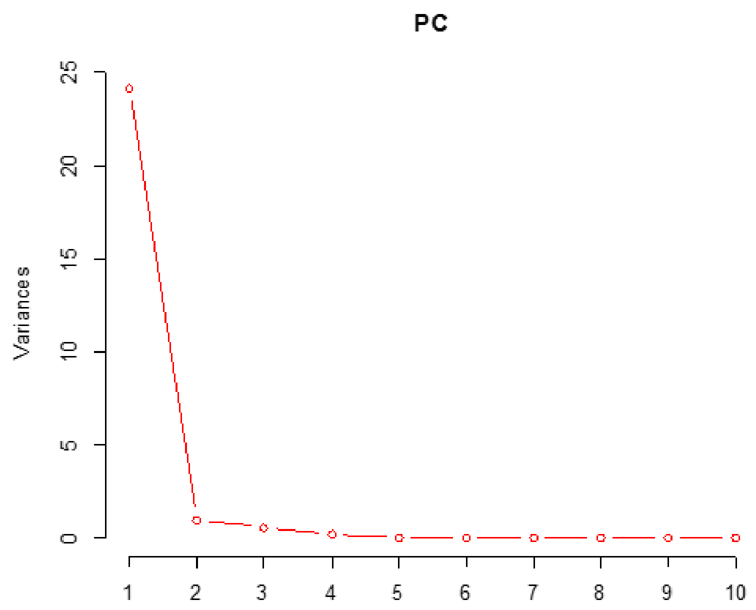
$$cor(\mathbf{H},\tilde{\mathbf{Z}}) = \max_{\mathbf{P}_{\mathbf{H}},\mathbf{P}_{\tilde{\mathbf{Z}}}} \frac{\mathbf{P}_{\mathbf{H}}\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)\mathbf{P}_{\tilde{\mathbf{Z}}}}{\left[\mathbf{P}_{\mathbf{H}}\left(\mathbf{H}^t\mathbf{H}\right)\mathbf{P}_{\mathbf{H}}\right]^{1/2}\left[\mathbf{P}_{\tilde{\mathbf{Z}}}\left(\tilde{\mathbf{Z}}^t\tilde{\mathbf{Z}}\right)\mathbf{P}_{\tilde{\mathbf{Z}}}\right]^{1/2}}$$

$\mathbf{H}$ vector of model variables

$\tilde{\mathbf{Z}}$ the portion of $\mathbf{Z}$ in to response set

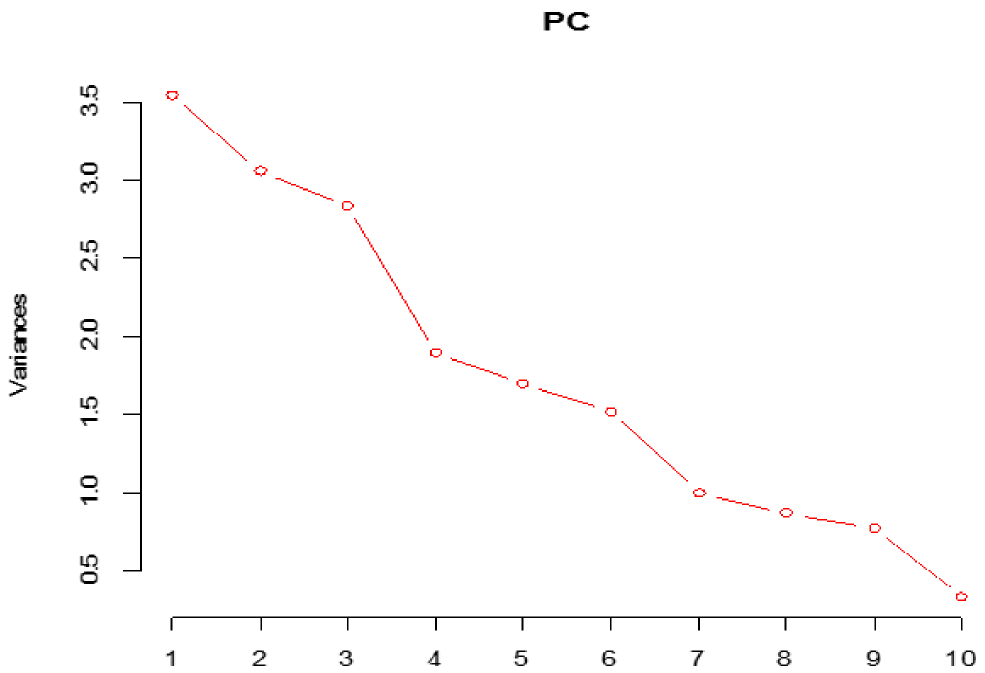
Simulation

- The use of PCs is illustrated in two simulation studies:
- Study 1 with auxiliary variables having strong correlation structure as the scree plot below illustrates



Simulation-cont.

- Study 2 where auxiliary variables are weakly correlated as the scree plot illustrates



Results

Table 1: LC on original population auxiliary variables vs. LC on population PCs –Study1

Sample size	Properties	Estimators	
		L. Calibration on X	L. Calibration on PCs
300	Rel.bias(%)	5.474	1.296
	S.E.	3519	935
	RMSE	8661	2094
600	Rel.Bias(%)	3.974	1.149
	S.E.	3135	846
	RMSE	6544	1864

Table 2: LC on original Sample auxiliary variables vs. LC on Sample PCs –Study1

Sample size	Properties	Estimators	
		L. Calibration on X	L. Calibration on PCs
300	Rel.bias(%)	3.930	0.192
	S.E.	21,936	11,202
	RMSE	22,660	11,206
600	Rel.Bias(%)	2.951	0.369
	S.E.	12,422	7608
	RMSE	13,134	7626

Table 3: LC on original population auxiliary variables vs. LC on population PCs-Study 2

Sample size	Properties	Estimators	
		L. Calibration on X	L. Calibration on PCs
300	Rel.bias(%)	0.033	0.010
	S.E.	5767	5588
	RMSE	5671	5588
600	Rel.Bias(%)	0.009	0.025
	S.E.	3769	3947
	RMSE	3769	3950

Table 4: LC on original Sample auxiliary variables vs. LC on Sample PCs-Study 2

Sample size	Properties	Estimators	
		L. Calibration on X	L. Calibration on PCs
300	Rel.bias(%)	0.007	0.006
	S.E.	5733	5645
	RMSE	5733	5645
600	Rel.Bias(%)	0.024	0.026
	S.E.	3913	3996
	RMSE	3914	4000

Table 5: PS on original population auxiliary variables vs PS on population PCs-Study 1

Sample size	Properties	Estimators			
		PS on X	Time (in hr)	PS on PCs	Time (in hr)
300	Rel.bias(%)	0.280	7	0.153	0.35
	S.E.	16,182		15,912	
	RMSE	16,188		15,914	
600	Rel.Bias(%)	0.169	36	0.264	1.30
	S.E.	10,899		10,757	
	RMSE	10,902		10,764	

Table 6: PS on original Sample auxiliary variables vs PS on Sample PCs-Study 1

Sample size	Properties	Estimators			
		PS on X	Time (in hr)	PS on PCs	Time (in hr)
300	Rel.bias(%)	0.255	0.25	0.125	0.18
	S.E.	16,161		16,010	
	RMSE	16,166		16,011	
600	Rel.Bias(%)	0.191	0.50	0.263	0.25
	S.E.	10,880		10,795	
	RMSE	10,884		10,801	

Table 7: PS on original population auxiliary variables vs. PS on population PCs-Study 2

Sample size	Properties	Estimators	
		PS on X	PS on PCs
300	Rel.bias(%)	0.587	0.670
	S.E.	18,522	19,738
	RMSE	18,894	20,212
600	Rel.Bias(%)	0.106	0.140
	S.E.	6035	6781
	RMSE	6072	6839

Table 8: PS on original Sample auxiliary variables vs. PS on Sample PCs-Study 2

Sample size	Properties	Estimators	
		PS on X	PS on PCs
300	Rel.bias(%)	0.196	0.264
	S.E.	10452	12222
	RMSE	10526	12334
600	Rel.Bias(%)	0.002	0.006
	S.E.	4155	4167
	RMSE	4155	4167

Table 9: Estimated model coefficients (Population auxiliary information-Study 2)

True Coefficients	δ_0 δ_1 δ_2 **1.306, -0.020, -0.083**					
Sample size	PS on X			PS on PCs		
	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$
300	1.128 (1.478)	-0.017 (0.000)	-0.036 (0.062)	1.182 (1.413)	-0.019 (0.000)	-0.039 (0.059)
600	1.205 (0.578)	-0.018 (0.000)	-0.066 (0.024)	1.238 (0.656)	-0.018 (0.000)	-0.069 (0.030)

Conclusion

- The results suggest the use of PC to be effective as this does not distort the results
- PCs are effective than original auxiliary variables in the conditions of the study in terms of computational effort.
- When the correlation structure is strong PCs are effective in PS calibration scheme than in LC scheme while weak correlation structure in auxiliary variables turns PCs more effective in LC than in PS

Thank you very much for your attention!