交通大数据

主成分分析

- 徐铖铖、李豪杰、郭延永
- guoyanyong@seu.edu.cn

- This lecture presents tools for illuminating ***structure in data*** in the presence of measurement difficulties, endogeneity, and unobservable or latent variables.
- Structure in data refers to relationships between variables in the data, including direct or causal relationships, indirect or mediated relationships, associations, and the role of errors of measurement in the models
- There are several approaches to uncovering data structure:
  - Principal components analysis is widely used as an exploratory method for revealing structure in data.
  - Factor analysis, a close relative of principal components analysis, is a statistical approach for examining the underlying structure in multivariate data.

Principal Components Analysis

- Principal components analysis has two primary objectives: to reduce a relatively large multivariate data set, and to interpret data.
- Principal components analysis “explains” the variance–covariance structure using a few linear combinations of the originally measured variables.
- Through this process a more parsimonious description of the data is provided——reducing or explaining the variance of many variables with fewer, well-chosen combinations of variables.

Principal Components Analysis

- If a large proportion (70 to 90%) of the total population variance is attributed to a few uncorrelated principal components, then these components can replace the original variables without much loss of information and also describe different dimensions in the data.
- Principal components analysis relies on the correlation matrix of variables, so the method is suitable for variables measured on the interval and ratio scales.

Principal Components Analysis

- If the original variables are uncorrelated, then principal components analysis accomplishes nothing.
- ✅ Observational data containing a large number of correlated variables
- ⚠ Experimental data with randomized treatments
- If it is found that the variance in 20 or 30 original variables is described adequately with four or five principal components (dimensions), then principal components analysis will have succeeded.

Principal Components Analysis

- Principal components analysis begins by noting that *n* observations, each with *p* variables or measurements upon them, is expressed in an $n \times p$ matrix ***X***:

$$X_{n\times p} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{n1} \end{bmatrix} \qquad (4.1)$$

- Principal components analysis is not a statistical model, and there is no distinction between dependent and independent variables.

Principal Components Analysis

- If the principal components analysis is useful, there are $K < n$ principal components,
- 寻求原指标的线性组合F_i。

$$
\begin{aligned}
F_1 &= u_{11}X_1 + u_{21}X_2 + \cdots + u_{p1}X_p \\
F_2 &= u_{12}X_1 + u_{22}X_2 + \cdots + u_{p2}X_p \\
&\cdots\cdots \\
F_p &= u_{1p}X_1 + u_{2p}X_2 + \cdots + u_{pp}X_p
\end{aligned}
$$

满足如下的条件：

1.每个主成分的系数平方和为1。即

$$u_{1i}^2+u_{2i}^2+\cdots+u_{pi}^2=1$$

2.主成分之间相互独立，即无重叠的信息。即

$$Cov\ (F_i,\ F_j)=0,\ i\neq j,\ i,\ j=1,\ 2,\ \cdots,\ p$$

3.主成分的方差一次递减，重要性依次递减，即

$$Var\ (F_1)\geq Var(F_2)\geq\cdots\geq Var(F_p)$$

F_1，F_2，…，F_k分别称为原变量的第一、第二、…、第p个主成分。

总体主成分的求解

- 矩阵知识回顾：

（1）特征根与特征向量

A、若对任意的k阶方阵C，有数字λ 与向量ξ 满足：

$\lambda\xi=C\xi$，则称 λ 为C的特征根，ξ 为C的相应于λ 的特征向量。

B、同时，方阵C的特征根 λ是k阶方程$|C-\lambda I|=0$的根。

（2）任一k阶方阵C的特征根 λ_j的性质：

$\sum_{j=1}^{k}\lambda_j=tr(C)=$矩阵C对角线上的元素之和

（3）任一k阶的实对称矩阵C的性质：

A、实对称矩阵C的非零特征根的数目=C的秩

B、k阶的实对称矩阵存在k个**实特征根**

C、实对称矩阵的不同特征根的特征向量是**正交的**

D、若 ξ_j是实对称矩阵C的**单位特征向量**，则

$$\xi_j' C \xi_j = \lambda_j$$

若矩阵 ξ，是由特征向量ξ_j所构成的，则有：

$$\xi_j' C \xi_j = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix}$$

主成分分析的目标：

1. 从相关的X_1，X_2，…，X_k，求出相互独立的新综合变量（主成分）Y_1，Y_2，…，Y_k。

2. $Y=(Y_1, Y_2, \dots, Y_k)'$ 所反映信息的含量无遗漏或损失的指标——方差，等于$X=(X_1, X_2, \dots, X_k)'$ 的方差。

X与Y之间的计算关系是：

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix} \quad \text{即} Y=AX$$

如何求解主成分？

从协方差矩阵出发求解主成分

（一）第一主成分：

设X的协方差矩阵为

$$\Sigma_X = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1P} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2P} \\ \vdots & \vdots & & \vdots \\ \sigma_{P1} & \sigma_{P2} & \cdots & \sigma_{PP} \end{bmatrix}$$

由于 $\sum_x$ 为非负定的对称阵，则有利用线性代数的知识可得，必存在正交阵U，使得

$$U'\Sigma_X U = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{bmatrix}$$

- 其中$\lambda_1,\lambda_2,\ldots,\lambda_p$为$\sum_x$的特征根，不妨假设

$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$

而U恰好是由特征根相对应的特征向量所组成的正交阵

$$\mathbf{U}=(\mathbf{u}_1,\cdots,\mathbf{u}_p)=\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

$$\mathbf{U}_i=\left(u_{1i},\ u_{2i},\cdots,\ u_{pi}\right)'\ i=1,2,\cdots,P$$

- 下面我们来看，是否由U的第一列元素所构成为原始变量的线性组合是否有最大的方差。

- 证明：设有P维正交向量 $\mathbf{a}_1=(a_{11},a_{21},\cdots,a_{p1})'$

$$F_1=a_{11}X_1+\cdots+a_{p1}X_p=\mathbf{a}'\mathbf{X}$$

$$V(F_1)=\mathbf{a}_1'\Sigma\mathbf{a}_1=\mathbf{a}_1'\mathbf{U}\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p\end{bmatrix}\mathbf{U}'\mathbf{a}_1$$

$$=\mathbf{a}_1'\left[\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_p\right]\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p\end{bmatrix}\begin{bmatrix}\mathbf{u}_1'\\ \mathbf{u}_2'\\ \vdots\\ \mathbf{u}_p'\end{bmatrix}\mathbf{a}_1$$

$$= \sum_{i=1}^{p} \lambda_i \mathbf{a}'\mathbf{u}_i\mathbf{u}_i'\mathbf{a} = \sum_{i=1}^{p} \lambda_i (\mathbf{a}'\mathbf{u}_i)^2$$

$$\le \lambda_1 \sum_{i=1}^{p} (\mathbf{a}'\mathbf{u}_i)^2 = \lambda_1 \sum_{i=1}^{p} \mathbf{a}'\mathbf{u}_i\mathbf{u}_i'\mathbf{a} = \lambda_1 \mathbf{a}'\mathbf{U}\mathbf{U}'\mathbf{a} = \lambda_1 \mathbf{a}'\mathbf{a} = \lambda_1$$

- 当且仅当 $a_1=u_1$时，即$F_1 = u_{11}X_1 + \ldots + u_{p1}X_p$时，有最大的方差 λ_1。因为 $\mathrm{Var}(F_1)=U'_1\Sigma_x U_1=\lambda_1$。
- 如果第一主成分的信息不够，则需要寻找第二主成分。

（二）第二主成分

在约束条件$\text{cov}(F_1, F_2)=0$下，寻找第二主成分

$$F_2 = u_{12}X_1 + \cdots + u_{p2}X_p$$

因为 $$\text{cov}(F_1, F_2) = \text{cov}(u_1'x, u_2'x) = u_2'\Sigma u_1 = \lambda_1 u_2' u_1 = 0$$

所以 $$u_2' u_1 = 0$$

则，对p维向量u_2，有

$$\begin{aligned} V(F_2) = u_2'\Sigma u_2 &= \sum_{i=1}^{p} \lambda_i \mathbf{u}_2'\mathbf{u}_i\mathbf{u}_i'\mathbf{u}_2 = \sum_{i=1}^{p} \lambda_i (\mathbf{u}_2'\mathbf{u}_i)^2 \\ &\le \lambda_2 \sum_{i=2}^{p} (\mathbf{u}_2'\mathbf{u}_i)^2 = \lambda_2 \sum_{i=1}^{p} \mathbf{u}_2'\mathbf{u}_i\mathbf{u}_i'\mathbf{u}_2 \\ &= \lambda_2 \mathbf{u}_2'\mathbf{U}\mathbf{U}'\mathbf{u}_2 = \lambda_2 \mathbf{u}_2'\mathbf{u}_2 = \lambda_2 \end{aligned}$$

2019/5/5

- 所以如果取线性变换： $F_2 = u_{12}X_1 + u_{22}X_2 + \cdots + u_{p2}X_p$

则F_2的方差次大

类推

$$
\begin{aligned}
F_1 &= u_{11}X_1 + u_{21}X_2 + \cdots + u_{p1}X_p \\
F_2 &= u_{12}X_1 + u_{22}X_2 + \cdots + u_{p2}X_p \\
&\cdots\cdots \\
F_p &= u_{1p}X_1 + u_{2p}X_2 + \cdots + u_{pp}X_p
\end{aligned}
$$

- 写为矩阵形式：

$$\mathbf{F} = \mathbf{U}'\mathbf{X}$$

$$\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_p) = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

$$\mathbf{X} = (X_1, X_2, \cdots, X_p)'$$

- 例1：设$x=(x_1,x_2,x_3)$的协方差矩阵为：

$$\Sigma=\begin{bmatrix}1 & -2 & 0\\ -2 & 5 & 0\\ 0 & 0 & 2\end{bmatrix}$$

从协方差矩阵出发，求解主成分

（1）求协方差矩阵的特征根

依据$|\Sigma-\lambda I|=0$ 求解

$$|\Sigma-\lambda I|=\begin{vmatrix}1-\lambda & -2 & 0\\ -2 & 5-\lambda & 0\\ 0 & 0 & 2-\lambda\end{vmatrix}=(1-\lambda)(5-\lambda)(2-\lambda)-(-2)(-2)(2-\lambda)=0$$

$$\lambda_1=5.83\quad \lambda_2=2\quad \lambda_3=0.17$$

（2）求特征根对应的特征向量

$$u_1=\begin{bmatrix}0.383\\ -0.924\\ 0.000\end{bmatrix}\quad u_2=\begin{bmatrix}0\\ 0\\ 1\end{bmatrix}\quad u_3=\begin{bmatrix}0.924\\ 0.383\\ 0.000\end{bmatrix}$$

（3）主成分

$$F_1 = 0.383x_1 - 0.924x_2$$
$$F_2 = x_3$$
$$F_3 = 0.924x_1 + 0.383x_2$$

（4）各主成分的贡献率及累计贡献率

第一主成分贡献率：$5.83/(5.83+2+0.17)=0.72875$

第二主成分贡献率：$2/(5.83+2+0.17)=0.25$

第三主成分贡献率：$0.17/(5.83+2+0.17)=0.02125$

第一和第二主成分的累计贡献率：

$$(5.83+2)/(5.83+2+0.17)=0.97875$$

由此可将以前三元的问题降维为两维问题。第一和第二主成分包含了以前变量的绝大部分信息97.875%。

□ 从协方差矩阵出发求解主成分的步骤：

1. 求解各观测变量 $X_l=(x_{1l},x_{2l},\ldots,x_{pl})'(l=1,2,\ldots,n)$ 的协方差矩阵。

2. 由X的协方差矩阵Σ，求出其特征根，即解方程 $|\Sigma-\lambda I|=0$ 可得特征根 $\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_p\geq0$

3. 求解 $\Sigma u_i=\lambda_i u_i$ 可得各特征根对应的特征向量 $U_1,U_2,\ldots,U_p$。

其中最大特征根的特征向量对应第一主成分的系数向量；第二大特征根对应的特征向量是第二大主成分的系数向量……

$$F_i=U_i'X,i=1,\cdots,k(k\leq p)$$

4. 计算累积贡献率，给出恰当的主成分个数。

5.对结果进行正确分析和合理解释。

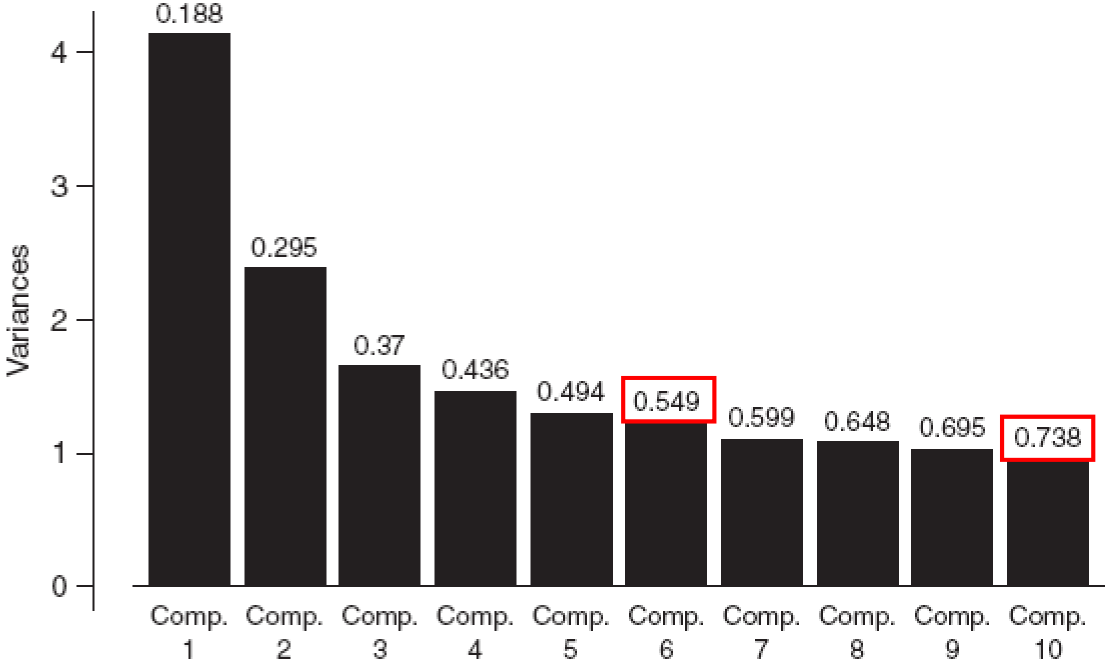
Variable Abbreviation	Variable Description
MODE	Usual mode of travel: 0 if drive alone, 1 if two person carpool, 2 if three or more person carpool, 3 if van pool, 4 if bus, 5 if bicycle or walk, 6 if motorcycle, 7 if other
HOVUSE	Have used HOV lanes: 1 if yes, 0 if no
HOVMODE	If used HOV lanes, what mode is most often used: 0 in a bus, 1 in two person carpool, 2 in three or more person carpool, 3 in van pool, 4 alone in vehicle, 5 on motorcycle
HOVDECLINE	Sometimes eligible for HOV lane use but do not use: 1 if yes, 0 if no
HOVDECREAS	Reason for not using HOV lanes when eligible: 0 if slower than regular lanes, 1 if too much trouble to change lanes, 2 if HOV lanes are not safe, 3 if traffic moves fast enough, 4 if forget to use HOV lanes, 5 if other
MODE1YR	Usual mode of travel 1 year ago: 0 if drive alone, 1 if two person carpool, 2 if three or more person carpool, 3 if van pool, 4 if bus, 5 if bicycle or walk, 6 if motorcycle, 7 if other
COM1YR	Commuted to work in Seattle a year ago: 1 if yes, 0 if no
FLEXSTAR	Have flexible work start times: 1 if yes, 0 if no
CHNGDEPTM	Changed departure times to work in the last year: 1 if yes, 0 if no
MINERLYWRK	On average, number of minutes leaving earlier for work relative to last year
MINLTWRK	On average, number of minutes leaving later for work relative to last year
DEPCHNGREAS	If changed departure times to work in the last year, reason: 0 if change in travel mode, 1 if increasing traffic congestion, 2 if change in work start time, 3 if presence of HOV lanes, 4 if change in residence, 5 if change in lifestyle, 6 if other
CHNGRTE	Changed route to work in the last year: 1 if yes, 0 if no
CHNGRTEREAS	If changed route to work in the last year, reason: 0 if change in travel mode, 1 if increasing traffic congestion, 2 if change in work start time, 3 if presence of HOV lanes, 4 if change in residence, 5 if change in lifestyle, 6 if other
I90CM	Usually commute to or from work on Interstate 90: 1 if yes, 0 if no
I90CMT1YR	Usually commuted to or from work on Interstate 90 last year: 1 if yes, 0 if no
HOVPST5	On your past five commutes to work, how often have you used HOV lanes
DAPST5	On your past five commutes to work, how often did you drive alone
CRPPST5	On your past five commutes to work, how often did you carpool with one other person
CRPPST52MR	On your past five commutes to work, how often did you carpool with two or more people
VNPPST5	On your past five commutes to work, how often did you take a van pool
BUSPST5	On your past five commutes to work, how often did you take a bus
NONMOTPST5	On your past five commutes to work, how often did you bicycle or walk
MOTPST5	On your past five commutes to work, how often did you take a motorcycle
OTHPST5	On your past five commutes to work, how often did you take a mode other than those listed in variables 18 through 24

(*continued*)

Variable Abbreviation	Variable Description
CHGRTEPST5	On your past five commutes to work, how often have you changed route or departure time
HOVSAVTIME	HOV lanes save all commuters time: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
HOVADUSE	Existing HOV lanes are being adequately used: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
HOVOPN	HOV lanes should be open to all traffic: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
GPTOHOV	Converting some regular lanes to HOV lanes is a good idea: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
GTTOHOV2	Converting some regular lanes to HOV lanes is a good idea only if it is done before traffic congestion becomes serious: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
GEND	Gender: 1 if male, 0 if female
AGE	Age in years: 0 if under 21, 1 if 22 to 30, 2 if 31 to 40, 3 if 41 to 50, 4 if 51 to 64, 5 if 65 or older
HHINCM	Annual household income (U.S. dollars): 0 if no income, 1 if 1 to 9,999, 2 if 10,000 to 19,999, 3 if 20,000 to 29,999, 4 if 30,000 to 39,999, 5 if 40,000 to 49,999, 6 if 50,000 to 74,999, 7 if 75,000 to 100,000, 8 if over 100,000
EDUC	Highest level of education: 0 if did not finish high school, 1 if high school, 2 if community college or trade school, 3 if college/university, 4 if post college graduate degree
FAMSIZ	Number of household members
NUMADLT	Number of adults in household (aged 16 or older)
NUMWRKS	Number of household members working outside the home
NUMCARS	Number of licensed motor vehicles in the household
ZIPWRK	Postal zip code of workplace
ZIPHM	Postal zip code of home
HOVCMNT	Type of survey comment left by respondent regarding opinions on HOV lanes: 0 if no comment on HOV lanes, 1 if comment not in favor of HOV lanes, 2 if comment positive toward HOV lanes but critical of HOV lane policies, 3 if comment positive toward HOV lanes, 4 if neutral HOV lane comment

Principal Components Analysis

Figure 8.1 shows a graph of the first ten principal components. The graph shows that the first principal component represents about 19% of the total variance, the second principal component an additional 10%, the third principal component about 8%, the fourth about 7%, and the remaining principal components about 5% each. Ten principal components account for about 74% of the variance, and six principal components account for about 55% of the variance contained in the 23 variables that were used in the principal components analysis. Thus, there is some evidence that some variables, at least, are explaining similar dimensions of the underlying phenomenon.



Principal Components Analysis

Table 8.2 shows the variable parameters for the six principal components. For example, the first principal component is given by

$$Z_1 = -0.380(HOVPST5) + 0.396(DAPST5) - 0.303(CRPPST5)$$

$$-\ 0.109(CRPPST52MR) - 0.161(BUSPST5) - 0.325(HOVSAVTIME)\ .$$

$$+\ 0.364(HOVOPN) - 0.339(GTTOHOV2) + 0.117(GEND)$$

Factor Loadings of Principal Components Analysis: HOV Lane Survey Data

Variable	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6
			Travel Behavior Variables			
HOVPST5	−0.380		−0.284	0.236		
DAPST5	0.396		0.274	−0.283		0.128
CRPPST5	−0.303		−0.223	0.240	0.282	0.221
CRPPST52MR	−0.109			0.167	0.196	−0.107
VNPPST5				−0.146		
BUSPST5	−0.161	−0.140	−0.227	0.112	−0.514	−0.395
NONMOTPST5						0.471
MOTPST5			0.104		0.381	−0.418
CHGRTEPST5					0.525	−0.302
			HOV Attitude Variables			
HOVSAVTIME	−0.325		0.301	−0.140		
HOVADUSE	−0.321		0.227	−0.133		
HOVOPN	0.364		−0.216	0.210		
GPTOHOV	−0.339	0.125	0.230	−0.115		
GTTOHOV2	−0.260		0.245	−0.153		
			Sociodemographic Variables			
GEND	0.117		0.388	0.180		−0.199
AGE			0.268	0.341	−0.363	−0.270
HHINCM		0.304	0.131	0.489		0.101
EDUC		0.188	0.247	0.443		0.247
FAMSIZ		0.429	−0.122			
NUMADLT		0.516	−0.188	−0.128	−0.133	
NUMWRKS		0.451	−0.242	−0.137		
NUMCARS		0.372	−0.106		0.107	−0.268

Note: Loadings < 0.10 shown as blanks.

Principal Components Analysis

All of the variables had estimated parameters (or loadings). However, parameters less than 0.1 were omitted from Table 8.2 because of their relatively small magnitude. The first principal component loaded strongly on travel behavior variables and HOV attitude variables. In addition, Z_1 increases with decreases in any non-drive-alone travel variables (HOV, Car Pool, Bus), increases with decreases in pro-HOV attitudes, and increases for males. By analyzing the principal components in this way, some of the relationships between variables are better understood.

$$Z_1 = -0.380(HOVPST5) + 0.396(DAPST5) - 0.303(CRPPST5)$$
$$- 0.109(CRPPST52MR) - 0.161(BUSPST5) - 0.325(HOVSAVTIME) .$$
$$+ 0.364(HOVOPN) - 0.339(GTTOHOV2) + 0.117(GEND)$$

or Loadings of Principal Components Analysis: HOV Lane Su

Variable	Comp. 1	Comp. 2	Comp. 3
			Travel Behavior Variables
HOVPST5	−0.380		−0.284
DAPST5	0.396		0.274
CRPPST5	−0.303		−0.223
CRPPST52MR	−0.109		
VNPPST5			
BUSPST5	−0.161	−0.140	−0.227
NONMOTPST5			
MOTPST5			0.104
CHGRTEPST5			
			HOV Attitude Variables
HOVSAVTIME	−0.325		0.301
HOVADUSE	−0.321		0.227
HOVOPN	0.364		−0.216
GPTOHOV	−0.339	0.125	0.230
GTTOHOV2	−0.260		0.245
			Sociodemographic Variables
GEND	0.117		0.388
AGE			0.268
HHINCM		0.304	0.131
EDUC		0.188	0.247
FAMSIZ		0.429	−0.122
NUMADLT		0.516	−0.188
NUMWRKS		0.451	−0.242
NUMCARS		0.372	−0.106

Note: Loadings < 0.10 shown as blanks.