基于电子设备大数据的人格、社会与行为变量相关研究

陈树铨 廖安迪 于明可 王亦凡

摘要 本文是基于MIT人类动力学实验室电子设备大数据集进行的数据探索与分析研究报告。首先我们对变量进行预处理，包括大五人格得分计算、R语言程序导出大数据和数据时间纵向次数统计等；其次我们对社会和人格变量分别做了因子分析，并对提炼出的有效因子做了不同分组依据的独立样本t检验、聚类分析和可视化呈现探索；最后我们对社会和人格的因子与行为变量进行了典型相关分析，寻找包含多指标的不同组间的相关关系。

关键词 大数据 预处理 因子分析 典型相关分析 人格

1.背景

本次数据探索研究基于由麻省理工学院人类动力学实验室（MIT Human Dynamics Lab）公布的“朋友与家庭（Friends and Family）”数据集。研究者通过手机应用程序的采集与定期问卷调查的方式，纵向收集北美某研究型大学一个年轻家庭居住区部分成员的多种数据，试图理解人们如何在不同社会层面的因素影响下做出决策，以及如何通过个人化或社会化的工具使人们做出很好的决策。

1.1 被试选取

研究人员Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal与Alex Pentland在某大学拥有超过400个年轻住户的居民区先后于2010年3月与2010年9月各选取55名、130名居民作为被试。这些被试全部已婚，且本人或配偶中至少一人供职于该校，约一半的家庭育有子女。被试从申请者中通过某种方式筛选，因而样本在社区与子社区中具有代表性。

1.2 数据收集

本数据集由四个部分构成。

第一部分为通过安卓手机应用程序每6分钟收集的传感器数据，包括由蓝牙信号获取的临近人员信息、由Wi-Fi信号获取的位置信息、通话与短信记录、手机应用程序的运行与安装、以及由加速器获取的活动信息。这些数据可用于描绘用户活动特点、关系网络、媒体消费与行为扩散模式。同时，有70%的被试同意记录其Facebook登录信息，用以获取其社交网络信息与线上沟通活动。

第二部分是由实验展开初期进行的问卷得到的个体人口学等信息。其中涵盖所属院系、族群、宗教信仰、住所区域与楼层，以及大五人格等社会和人格心理学指标、宿舍活动的参与度、收入与包括体育锻炼频率、心境与自信程度在内的生活方式信息。其中有部分数据未被公布。

第三部分是月度、周度与日度进行的自我报告问卷，揭示被试的社会活动与

社会互动状况。月度报告包括对于人际关系的自我知觉、群体关系、社会互动等；周度报告的主要变量为饮食方式、与谁一同就餐、各类娱乐方式的频率与伙伴、对手机应用程序的使用与评价、获取或提供包括照顾孩子等帮助的对象。此外，被试需要每日报告前一日的心情、睡眠等活动。

第四部分是被试自愿提供的收据、信用卡账单与社交网站等信息。但这部分数据的被试量较少，事实上难以进行分析。

需要强调的是，后三部分的数据是被试在手机上填写、通过网络收集的。

1.3 研究现状

目前，基于这一系列数据集的研究已发表 9 篇。研究的问题涉及面十分广泛，具体来说有以下五个研究，其他研究的问题跟这五个研究类似，因而主要以这五个研究为例予以介绍。

Aharony 等人最初报告了不同奖励形式激励人们进行体育锻炼的效果，同时比对社会互动与经济状况间的关系，以及被试的社会结构对人们决策习惯的影响。Yves-Alexandr 等研究者通过大五人格分布的正态性的检验，认为数据分布上适合做分析。在用不同信息解释人格的研究中，研究者尚未发现有一致解释力的指标，但通话时间和外向性显著相关，宜人性与孩子数量有较显著相关。Yaniv Altshule 等研究者基于电话、短信、见面等的数据，对参与本研究的被试进行了社会测量学分析，并根据分析结果制作出了社区成员心理距离连线图。Aniv Altshuler 等研究者用电话数据建模和预测社交行为变量，并对此做了正确度的检验。Sai T Moturu 等研究者挖掘了睡眠、情绪和社会关系之间的联系，发现睡眠质量跟情绪的正负性有很大的关系，而情绪的正负效价与社会融入程度又有较大的关系。

1.4 研究局限

尽管目前对于此数据集的研究已经很多，但纵观所有研究，存在的局限和可供用的空间还是比较大的：

第一，已有的研究没有考虑对数据进行分组因子分析和主成分分析，因而使得数据量较为冗余，没有提炼出有效而有力的的指标。

第二，数据预处理阶段没有考虑进行时间累加，比如通话总次数的统计、短信总数的统计。

第三，对于人格和行为变量指标的变量组间关系的研究仍处于空白状态，典型相关分析是一个可以尝试的方法。

总而言之，数据还有可继续利用的空间。

2.方法

本次研究采用 2010 年 3-5 月，参与被试 56 名。数据分析和研究方法步骤可

以简化如图 1 所示。

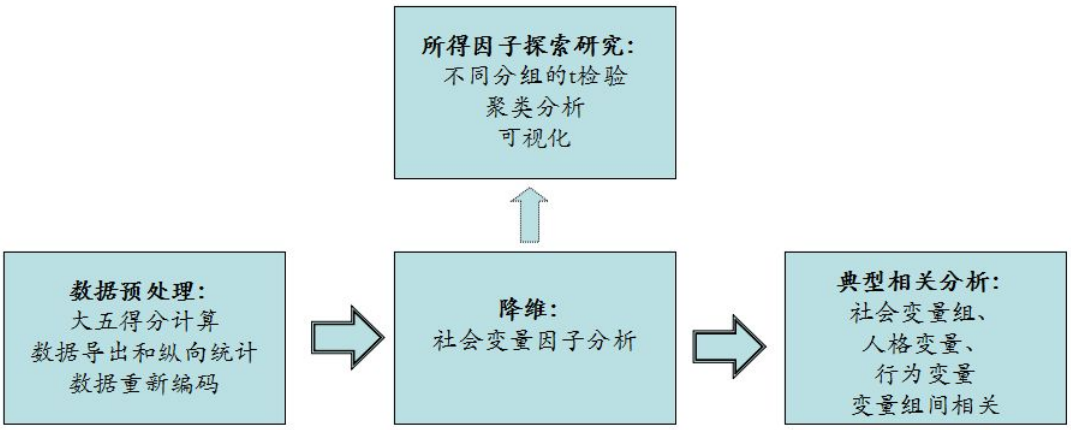


图 1 研究方法步骤图解

（实线框箭头为研究方法主要进路，虚线框箭头为辅助进路）

我们的研究主要包括三个步骤。

第一步是数据的预处理。我们根据大五人格问卷 44 题版的计算标准进行了个别题目的反向计分和各维度记分、标准化为百分制得分。与此同时，我们用 R 读取大数据，获得了通话、短信、电池状态等变量，并纵向统计以下变量：2010 年 3-5 月每个月的电话打进量、电话打出量、未接电话量、短信接收量、短信发出量。此外，对于以字符串编码的社会变量，我们进行了重新编码。

第二步是以因子分析为方法的降维。我们做了以下七组变量的因子分析以实现降维。一是通讯变量，与当地亲属、当地（指大波士顿地区）朋友与远方朋友六种沟通方式（陆话与电话、网络电话、短信、邮件以及网络聊天与即时消息，当地的还有见面）共 17 个关于频繁程度的变量。二是社区关系变量，包括感受到的融入感、期待的融入感、以及各种朋友在社区所占比例等 9 个变量。三是各种手机应用程序类别的喜爱程度与常用程度，分别用被试最喜爱或最常用的五个程序所属的类别表示，含游戏、娱乐、生活方式等 11 个类别。四是被试对获知手机应用程序消息的各类信息渠道的使用频率与价值评价，有 22 个变量，其中，信息渠道包括配偶、朋友等人以及电视广告、互联网搜索、Funf 论坛等媒介。五是对自我共情能力的评价，包括在快乐、悲伤、气愤、恐惧和紧张场合下对于自身情感的唤起，共 15 个变量。六是情绪与心境变量，包括两个部分，情绪变量包括在过去一周内感觉到激动的、低落的、内疚的等 32 种情绪的天数，而心境变量包括在过去一个月内感觉到开心满意、镇静平和等 5 种情绪的天数。七是生活方式变量，包括饮食、有氧运动、锻炼、压力、情感、休息和睡眠共 7 个变量。

第三步是典型相关分析。

3.结果

3.1 预处理

大五人格得分依次为外向性得分、宜人性得分、尽责性得分、神经质得分和开放性得分。

数据的导出和纵向统计得到了电话、短信、电池 2010 年 3-5 月三个月的信息。

至于重新编码，主要操作对象是 2010 年 4 月调查中被试向研究者报告的数项与社会关系有关的问题。统计的变量包括:

第一，被试分别与不同对象通过不同媒介进行通讯的频率。其中，联络的对象包括当地亲属、当地（指大波士顿地区）朋友与远方朋友，通讯方式涵盖陆话与电话、网络电话、短信、邮件以及网络聊天与即时消息；对于当地的亲属或朋友，通讯方式还包括见面。以上诸变量有六点可选，分别是从不、少于一月一次、一月一至两次、一周一至两次、大部分日子，以上 17 个变量按通讯的频率由低至高被编码为由 1 至 6。

第二，有关社区关系的自我报告：感受到的、及期待的社区融入感（1-一点都不，2，3-有点，4，5-非常）；在居住区的公共空间与邻居对话的频率（1-从不，2,-少于一月一次，3-一月一至两次，4-一周一至两次，5-大部分日子）；除自己公寓，认识的本楼层家庭数（0 至 9）；不同类别个体（亲密朋友；定期交往者；与之进行学术讨论者）中来自居住区者占据的比例（1-无(0%)，2-较少(1-20%)，3-一些(21-40%)，4-约一半 (41-60%)，5-大多(61-80%)，6-几乎全部(81-99%)，7-全部(100%)）；为不同类别个体（亲戚；居住区的朋友；非居住区的朋友；为非朋友提供付费服务）照看孩子的频率（1-从不，2-一年数次，3-一月一次，4-一月数次，5-一周一次，6-多于一周一次）。我们使用上述标注的方式对频次等变量用数字进行编码。通过初步的探索性分析，照看亲戚孩子与收费照看非朋友孩子两个变量由于斜度与峰度系数过高而被剔除。

第三，手机应用程序类别的使用频率与偏好。研究者要求被试分别填写自己最常用的与最喜爱的 5 个手机应用程序，并注明其它们各自所属的类别。被试可在游戏、娱乐、手机个人化、新闻与天气、运动、购物、社交、通讯、生活方式、效率与工具、其他这 11 个选项中进行选择，若选择其他，还需注明特定的类别名称。调查结果中，分别有被试单独标注交通、宗教两个程序类别。使用被试标明的类别名替换“其他”后，变量增加至 24 个。在预处理过程中，我们得到各类别中每个被试最喜爱或最常用的程序数目，最小值为 0，最大值为 5。手机个人化、运动、购物、交通这四个类别在最喜爱与最常用两个指标中的至少一项中

出现标准差为零的现象，这些变量与两个指标所对应的数据均被剔除。

第四，获取手机应用程序信息的来源的使用频率与价值评价。受调查者对获取新应用程序信息的不同来源的使用频率（1-从不，2-一月一次，3-几周一次，4-一周一次，5-一周数次，6-每天）与价值（1-完全无价值，2，3，4，5-非常有价值）的评价。被评价的信息来源包括被试的配偶、朋友、熟人、其他人、安卓应用市场、FunF 论坛、网络搜索、网络广告、新闻文章、其他媒体。使用上述所标注数字对数据进行编码。

第五，自我共情能力的内容：对于所有表述，均是按照 5 点量表（1-不符合，2, 3-中立, 4,5-符合）进行转换。

第六，情绪与心境的内容：对于所有表述，均是按照 7 点量表（1-非常不符合，2, 3, 4-中立, 5, 6, 7-非常符合）进行转换。

第七，生活方式的内容：饮食的健康程度（1-非常不健康, 2, 3, 4-平均水平, 5, 6, 7-非常健康）；应对压力的自如程度（1-非常不自如，2, 3, 4-平均水平, 5, 6, 7-非常自如）。其余变量是一周内实际天数或者一天内实际小时的度量，因此不做重新编码。

3.2 因子分析

3.2.1 社会关系之通讯

由于社会关系的原始维度较多，且各维度间存在冗余，下面通过相关矩阵对其进行因子分析，以达到降维的目的。因子分析存在取样不足的问题，变量间相关关系较弱（KMO=.548），Bartlett 球形检验显著（$p < .001$），同时只解释 65.273% 的方差。共提取 5 个因子作为衡量社会关系的维度。

表3.2.1 与不同对象通过不同媒介进行通讯的频率的因子分析旋转成分矩阵

	成份				
	使用skype通讯频率	与亲戚电信通讯或见面频率	与朋友电信通讯或见面频率	使用网络短信通讯频率	使用邮件通讯频率
当地亲戚陆话与电话	-.184	.718	.070	.034	.156
当地亲戚网络电话skype	.648	-.018	-.420	.298	-.132
当地亲戚短信	-.069	.748	.153	.171	-.114
当地亲戚邮件	-.077	.573	.000	.214	.500
当地亲戚网络短信	-.021	.347	-.214	.755	.173
当地朋友陆话与电话	.079	-.176	.707	-.028	.179
当地朋友网络电话skype	.749	-.181	.146	.089	.194
当地朋友短信	-.031	.313	.749	.100	.033
当地朋友邮件	-.023	.057	.167	.021	.793
当地朋友网络短信	.028	.000	.125	.839	-.026
远方朋友陆话与电话	.459	.454	.410	-.159	-.018
远方朋友网络电话skype	.837	-.004	.130	.016	-.030
远方朋友短信	.397	.389	.602	.268	.125
远方朋友邮件	.495	.149	-.024	.121	.629
远方朋友网络短信	.372	.013	.225	.708	.209
见亲戚	.330	.556	-.029	-.022	.223
见当地朋友	.029	.106	.459	.140	.538

其中，不同通讯方式表现出较大程度的差异，而使用传统的电话、短信与直接见面这三种方式又依据对象为亲人或朋友而有所不同。

将提取的五个变量按其解释情况依次命名为“使用 skype 通讯频率”、“与亲戚电信通讯或见面频率”、“与朋友电信通讯或见面频率”、“使用网络短信通讯频率”、“使用邮件通讯频率”。表述内容与其名称相一致。

3.2.2 社会关系之社区关系

以同样的方式对社区关系变量进行因子分析，变量间显示较强的相关关系（KMO=.777）且 Bartlett 球形检验显著（$p < .001$）。提取的三个因子共可解释 75.436%的变异。依据其对各原始变量解释程度不同，将其命名为“社区融入度”、“帮助照看孩子频率”与“社区熟识度”。“社区融入度”衡量被试的社区融入感与社区中熟人的比例；“帮助照看孩子频率”表述的意义与名称一致；“社区熟识度”表示被试与邻居的熟悉程度。

表3.2.2 社区关系变量的因子分析旋转成分矩阵

	成份		
	社区融入度	帮助照看孩子频率	社区熟识度
实际的社区融入感	.812	.071	.283
期待的社区融入感	.833	-.037	.152
和邻居对话	.193	.345	.766
认识同层的家庭数	.132	.017	.862
亲密朋友住在w的比例	.821	.225	-.015
定期交往的人住在w的比例	.825	.244	.115
你和本社区多少人讨论专业生活	.736	.436	.219
照看w朋友孩子	.306	.787	.196
照看非w朋友小孩	.058	.894	.086

3.2.3 手机应用程序类别的使用与偏好

在剔除手机个人化、购物、运动、交通这四个类别后，仍剩余 8 个程序类别。下面使用相关矩阵进行因子分析，对最常用、最喜爱的 8 种类型共 16 个变量同时降维。结果显示，变量间相关关系较弱（KMO=.492），但 Bartlett 球形检验显著（$p < .001$），得到 6 个因子，能够解释 80.084%的变异。尽管案例数目较少，但仍能发现人们对于同类型应用程序的喜爱程度与常用程度大体一致，揭示其态度与行为的一致性。

其后，按最常用的与最喜爱的将变量分为两组，分别进行因子分析。结果显示，生活方式与宗教两类应用的成分表达始终保持一致，因此将后者并入前者。

表3.2.3 获取应用程序信息来源的使用频率与价值评的因子分析旋转成分矩阵

	成份			
	生活方式-娱乐	效率工具-社交平台	游戏-通讯	新闻天气
UsdGames	-.009	-.140	.868	-.005
UsdEntert	.763	.097	.198	-.139
UsdNwsWth	-.005	.046	-.020	.837
UsdSocial	-.052	-.658	-.353	.140
UsdCommu	.281	-.337	-.473	-.315
UsdLifsty	-.675	.047	.040	-.331
UsdPrdTol	.049	.825	-.193	.158
EnjGames	.222	-.031	.825	-.004
EnjEntert	.750	.134	-.059	-.115
EnjNwsWth	.138	.061	.154	.825
EnjSocial	-.035	-.535	-.302	.369
EnjCommu	.388	-.223	-.445	-.196
EnjLifsty	-.746	.017	-.003	-.415
EnjPrdTol	.046	.778	-.180	.185

尽管样本容量较小，变量间相关较低（KMO=.460），但在对余下 7 个类别应用程序的最喜爱与最常用应用数目总计 14 个变量同时进行因子分析时，Bartlett 球形检验显著（$p < .001$），有 63.905%的差异得到解释。应用程序的类别在最喜欢与最常用这两项上均表现出高度一致。

表 3.2.3 中，变量名中的“Usd”指最常使用，“Enj”表示最喜欢；第四个字母后的内容表示程序的类别，依次为游戏、娱乐、新闻天气、社交工具、通讯、生活方式和效率工具。

最终提取的 4 个因子依据取其解释情况分别被命名为“生活方式-娱乐”、“效率工具-社交平台”、“游戏-通讯”、“新闻天气”四个类别，作为解释人们对应用程序的使用习惯与态度的四个分类维度。人们在生活方式与娱乐、效率工具与社交平台、游戏与通讯这几对类别的打分上表现出一定趋势——喜爱并常用娱乐类程序者，较少使用生活方式类程序；喜爱效率工具者更少使用社交平台；多用游戏的被试更少地使用通讯软件等。

3.2.4 获取手机应用程序信息的来源的使用频率与价值评价

以相似的方式对被试对手机应用程序的不同信息来源的使用频率和价值评价（共计 22 个变量）进行因子分析。其中，变量间相关关系不够强（KMO=.566），Bartlett 球形检验显著（$p < .001$），提取出 4 个特征值大于 1 的因子，能够解释 68.843%的差异。

表3.2.4 获取应用程序信息来源的使用频率与价值评的因子分析旋转成分矩阵

	成份			
	中信任源使用	高信任源	中信任源评价	低信任源
RtSpouse	.446	.722	.023	-.030
RtFriend	.472	.624	.241	.086
RtAcquai	.036	.710	.050	.234
RtOthPer	.680	.236	-.008	.188
RtAdrMar	.070	.210	.244	.613
RtFForum	.246	.569	.058	.415
RtItnSch	.497	-.121	.002	.714
RtItnAds	.873	.053	.179	.140
RtNwsAtc	.854	.084	.117	.273
RtTVads	.897	.040	.292	-.027
RtOther	.858	-.138	.239	-.013
VlSpouse	-.073	.748	.092	-.297
VlFriend	-.119	.639	.506	.013
VlAcquai	-.091	.666	.559	.094
VlOthPer	.120	.381	.735	.029
VlAdrMar	-.225	.419	.401	.563
VlFForum	-.022	.561	.064	.450
VlItnSch	.176	-.051	.140	.791
VlItnAds	.373	.153	.722	.318
VlNwsAtc	.374	.184	.485	.325
VlTVads	.343	.081	.801	.062
VlOther	.147	-.042	.793	.181

表 3.2.4 中的变量名称中，Rt 表示使用频率，Vl 表示价值评价；第三个字母开始的部分表示信息来源，依次为配偶、朋友、熟人、其他人、安卓应用市场、Funf 论坛、网络搜索、网络广告、新闻文章、电视广告、其他媒体。

通过比对旋转后的成分矩阵发现，不同类别的信息来源在一定程度上在使用频率与个人的主观评价间存在关联，且这些信息来源可进一步被分为不同的类别。根据每个因子最大程度体现的变量的特点，我们将其命名为“中信任源使用”、“高信任源”、“中信任源评价”、“低信任源”。

其中，“低信任源”指安卓应用市场、互联网搜索这类信息来源；“中信任源”指被试不熟悉的人、互联网广告、新闻文章、电视广告及其他媒体；而“高信任源”则是指配偶、朋友、熟人与 Funf 论坛这类比较熟悉或是比较值得信赖的信息来源。

查看降维后的数据可以验证，被试高、低信任源在使用与评价上具有高度的一致，对其价值评价更高的被试也倾向于更多地使用它们。然而，对于中信任源的使用频率与价值评价出现分离，需要使用两个维度进行表示。这说明人们对不熟悉的人、新闻媒体等处得到的信息的评价与实际使用的总体分布存在差异，有些人倾向于对这类来源使用较多而评价较低，但另一些人对此使用较少而评价较

高。

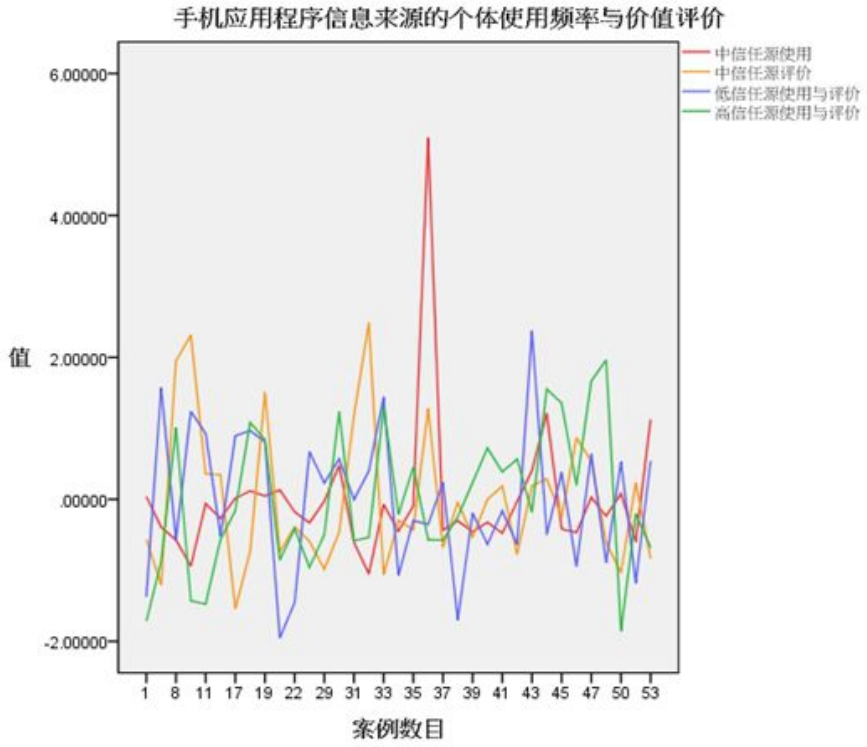


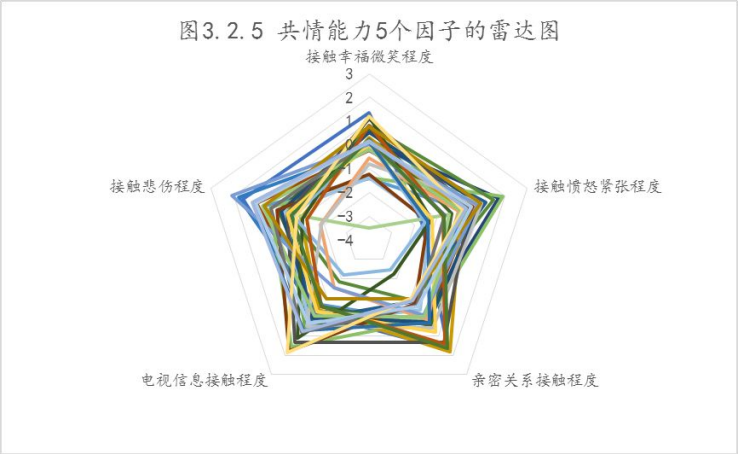
图 2 各被试对不同类别信息来源的使用频率与价值评价折线图

3.2.5 自我共情能力

共情能力的度量，从原始的 14 个维度中进行降维，根据因子分析的结果，提取了 5 个因子作为 mood 的度量，共解释了原始数据 78.3%的方差，并分别将这 5 个因子命名为接触幸福微笑程度，接触愤怒紧张程度，亲密关系接触程度，电视信息接触程度和接触悲伤程度。

表 3.2.5 自我共情能力旋转成分矩阵

	幸福微笑	愤怒紧张	亲密关系	电视信息	悲伤
Someone I'm talking with begins to cry	.067	.218	.206	.427	.586
Being with a happy person picks me up	.895	.009	.091	-.096	-.099
Someone smiles warmly at me	.709	.298	.279	.279	.085
People talk about the death of their loved ones	.127	.204	.119	.816	.116
See the angry faces on the news.	-.263	-.035	-.257	.736	.047
Look into the eyes of the one I love	.261	.127	.751	.100	.074
Be around angry people	.337	.829	-.010	.099	.016
Watching the fearful faces of victims on the news	.418	.288	.112	.731	.024
The one I love holds me close.	.261	.173	.875	-.134	.107
Overhearing an angry quarrel.	.213	.789	.124	.160	.079
Being around happy people	.797	-.001	.271	.170	.177
The one I love touches me.	.684	.204	.516	-.110	-.047
Around people who are stressed out.	-.235	.853	.194	.185	.104
Cry at sad movies.	.027	.008	.004	.038	.866
Listening to the shrill screams of a terrified child	-.112	.442	-.552	-.187	.450

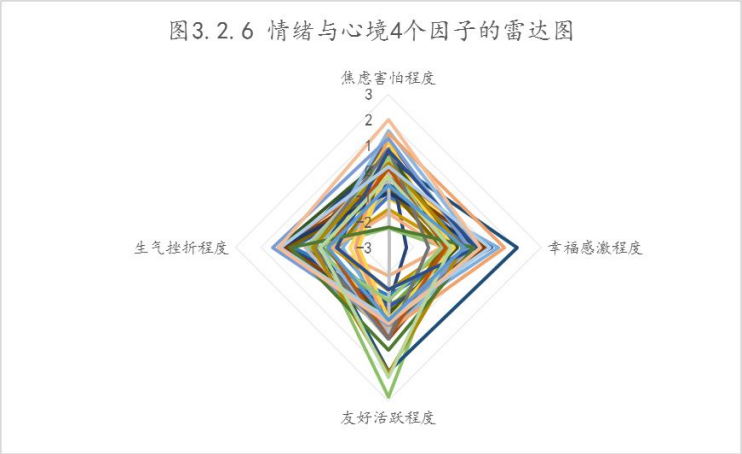


3.2.6 情绪与心境

关于情绪，由于原始维度较多，共有 37 项表述，需要进行降维处理。根据因子分析的结果，提取了 4 个因子作为情绪与心境的度量，解释了原始数据的 79.6%的方差，分别将这 4 个因子命名为焦虑害怕程度，幸福感激程度，友好活跃程度，生气挫折程度。

表 3.2.6 情绪与心境旋转成分矩阵

	焦虑害怕程度	幸福感激程度	友好活跃程度	生气挫折程度
Interested	-.282	.211	-.211	-.018
Distressed	.473	-.176	.043	-.008
Excited	-.378	.192	-.118	-.004
Upset	.549	-.320	.114	-.101
Strong	-.019	.427	-.115	-.069
Guilty	.748	-.009	.099	-.022
Scared	.700	-.247	.056	-.115
Hostile	.512	.065	.006	-.004
Enthusiastic	-.379	.103	-.080	-.211
Proud	-.384	.148	-.078	.093
Irritable	.456	-.340	.214	-.157
Alert	-.051	.426	-.163	.051
Ashamed	.582	.008	.042	.001
Inspired	-.179	.053	-.090	.186
Nervous	.520	-.086	.077	-.058
Determined	.025	.478	-.060	-.030
Attentive	-.104	.495	-.049	.106
Jittery	.361	.204	.115	-.238
Active	-.140	.610	-.069	-.071
Afraid	.667	-.218	-.138	-.033
Impatient	.213	-.322	-.051	.013
Disgusted	.425	-.250	-.140	.082
Enjoying myself	-.411	.003	-.316	.110
Depressed/blue	.589	-.131	.010	-.098
Competent/capable	-.458	.257	-.084	-.001
Worried/anxious	.517	-.045	.004	-.033
Criticized/putdown	.463	.009	-.053	.024
Grateful	-.100	.164	-.259	.022
Happy	-.245	.130	-.257	.172
Warm/friendly	-.303	.531	-.167	-.077
Guilty/regretful	.671	-.038	.032	-.030
Compassionate	-.008	-.153	-.041	.227
开心满意天数/月	-.089	-.133	-.405	.882
镇静平和天数/月	.077	.565	.233	.783
焦虑紧张天数/月	.278	-.791	.490	.025
伤心沮丧天数/月	.076	-.306	.725	-.167
生气挫折天数/月	-.372	-.151	.807	-.084



3.2.7 生活方式

生活方式方面的变量，由于原始维度较多，共有 9 项表述，对此进行因子分析。因为 Initial Survey 几乎没有被试数据缺失，因此选用 Initial Survey 的 7 个变量数据进行降维，而 Initial Survey-2010-10-continue 主要用于时间序列的比较。因子分析完成后，共提取了 3 个因子作为生活方式的度量，解释了原始数据 67.7% 的方差，并将其命名为体育锻炼程度，睡眠休息程度，饮食与压力应对程度。从时间序列的对比图可知，理想很丰满，现实很骨感。

表 3.2.7 生活方式的旋转成分矩阵

	体育锻炼程度	睡眠休息程度	饮食与压力应对程度
健康饮食实际1	.004	-.008	.813
有氧运动实际1	.918	.028	.094
锻炼实际1	.909	.077	.019
压力应对1	.158	.060	.741
情感支持1	-.022	.205	.438
休息时长1	.080	.846	.204
睡眠实际1	.034	.902	.032

图3.2.7.1 生活方式3个因子的雷达图

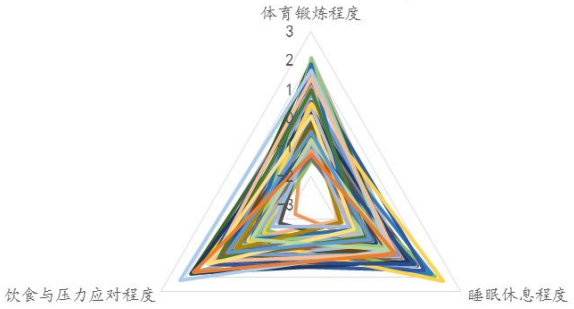


图3.2.7.2 饮食实际与理想对比雷达图

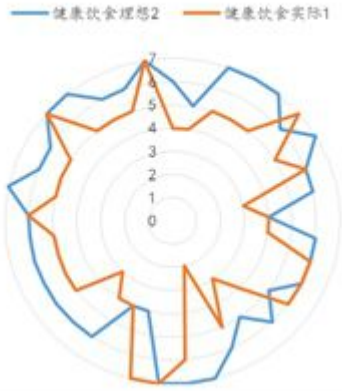
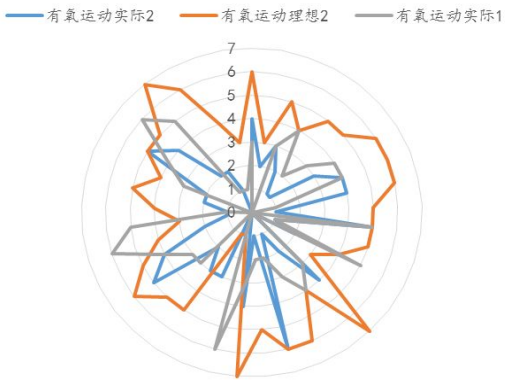
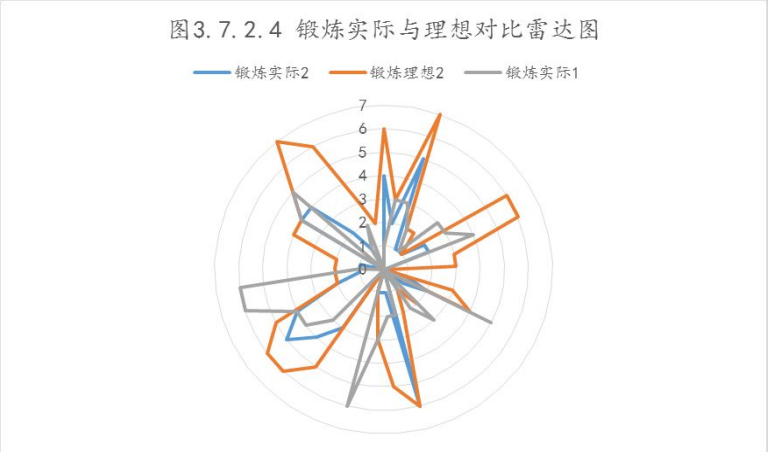
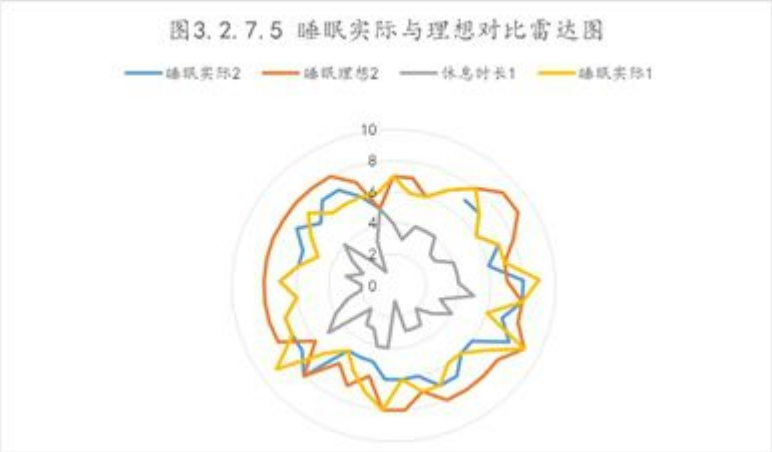
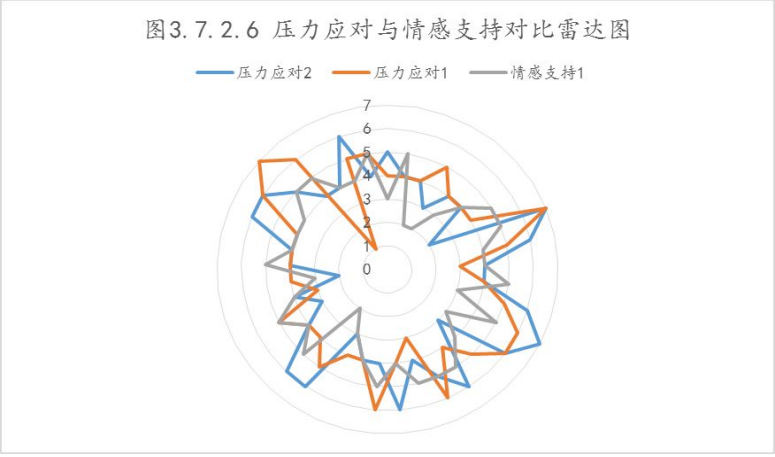


图3.2.7.3 有氧运动实际与理想对比雷达图









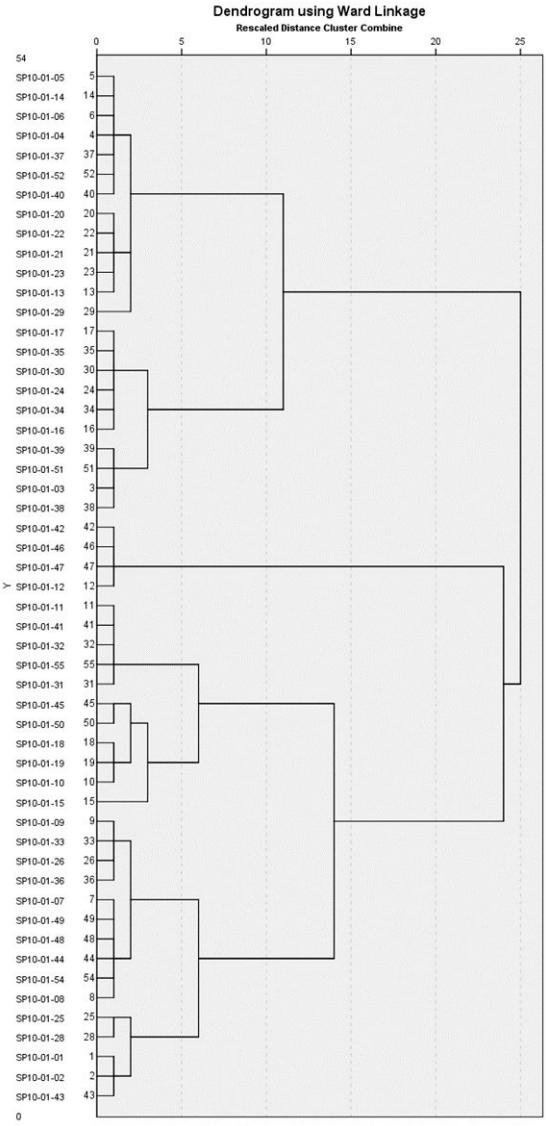
3.2.8 基于因子分析的探索性数据分析

3.2.8.1 夫妇行为及态度模式的聚类分析

在数据预处理及因子分析的基础之上，尝试分别用行为和态度变量进行聚类分析，观察夫妇的行为及态度模式是否存在较高的一致性。聚类分析采用的是层

级聚类法，类别间相似度度量为 Ward 法。

当使用社区融入度、帮助他人照看孩子频率和社区熟识度三个维度作为聚类指标时，可以发现，夫妇通常在前两次聚类就可聚为一类，即他们在社区融入度、帮助他人照看孩子频率和社区熟识度三个维度上行为与态度模式较为类似，说明孩子和社区和维系夫妇情感纽带的重要因素。但是，在其他维度的聚类分析上，均没有发现如此显著的分类变量。



3.2.8.2 性别差异的独立样本 t 检验

由于有着男性与女性在行为与态度模式上可能存在差异的猜想，因此采用了独立样本 t 检验的方法，对行为与态度变量进行了分析。结果发现女性的神经质得分显著高于男性，但男性的开放性与创造力得分显著高于女性；在接触悲伤的信息时，女性比男性更容易共情；而男性比女性更容易感觉到焦虑和害怕。

表 3.2.8.2a 关于性别在大五人格及创造力得分上的独立样本 t 检验

	Gender	Mean	T	Df	Sig.	Mean Difference	Std. Error Difference
外向性得分	Male	63.4615	-.721	51	.474	-3.39031	4.70179
	Female	66.8519					
宜人性得分	Male	74.1026	-1.609	51	.114	-4.58056	2.84656
	female	78.6831					
神经质得分	Male	51.4423	-2.209	51	.032	-8.55769	3.87433
	female	60.0000					
尽责性得分	Male	71.7949	-.342	51	.734	-1.12694	3.29331
	female	72.9218					
开放性得分	Male	76.1538	2.148	51	.036	7.93162	3.69246
	female	68.2222					
创造力得分	Male	60.3158	2.616	39	.013	8.67943	3.31795
	female	51.6364					

表 3.2.8.2b 关于性别在共情能力得分上的独立样本 t 检验

	Gender	Mean	T	Df	Sig.	Mean Difference	Std. Error Difference
接触幸福微笑程度	Male	.1387912	.793	29	.434	.28683507	.36164021
	female	-.1480439					
接触愤怒紧张程度	Male	.0740452	.420	29	.678	.15302674	.36443545
	female	-.0789815					
亲密关系接触程度	Male	.1838099	1.059	29	.298	.37987387	.35867072
	female	-.1960639					
电视信息接触程度	Male	.0536572	.304	29	.763	.11089159	.36496117
	female	-.0572344					
接触悲伤程度	Male	-.3537808	-2.154	29	.040	-.73114689	.33939217
	Female	.3773661					

3.2.8.3 预测孩子数量、焦虑害怕程度的多重回归分析

3.2.8.3a 预测孩子数量的多重回归模型

在人口学统计资料中，我们选取了孩子数量作为因变量，力图通过手机应用程序喜爱常用程度的行为变量与社区融入的社会变量来进行预测。多重回归模型结果建立之后，发现帮助他人照看孩子概率、社区熟识度和手机应用程序游戏减通讯喜爱常用程度具有较为有力的解释力度，并且当帮助他人照看孩子概率越

高，社区熟识度越高，手机应用程序游戏减通讯喜爱常用程度越低时，对应的样本更可能有孩子。

表 3.2.8.3a 预测孩子数量的多重回归模型

	B	Std. Error	t	Sig
常数	.490	.068	7.150	.000
社区融入度	.185	.070	2.653	.013
帮助他人照看孩子频率	.161	.067	2.394	.024
社区熟识度	.101	.089	1.138	.266
app生活方式减娱乐喜爱常用程度	-.065	.072	-.892	.381
app效率工具减社交平台喜爱常用程度	-.030	.070	-.425	.675
app游戏减通讯喜爱常用程度	-.170	.070	-2.429	.022
app新闻天气喜爱常用程度	.044	.072	.608	.548

3.2.8.3b 预测焦虑害怕程度的多重回归模型

在社会变量中，我们选取了焦虑害怕程度作为因变量，力图通过生活方式与共情能力这两个社会变量来进行预测。多重回归模型结果建立之后，发现在第一个模型中，睡眠休息程度和饮食与压力应对程度具有较为有力的解释力度，即睡眠休息程度越高，饮食与压力应对程度越高，则对应的被试的焦虑害怕程度越低；而在第二个模型中，亲密关系接触程度和电视信息接触程度具有较为有利的解释力度，即越与亲密关系的人接触多，越少暴露在电视信息之下，则焦虑害怕程度越低。

表 3.2.8.3b 预测焦虑害怕程度的多重回归模型

	B	Std. Error	T	Sig
常数	.032	.148	.214	.832
体育锻炼程度	.122	.137	.896	.377
睡眠休息程度	-.355	.152	-2.336	.026
饮食与压力应对程度	-.316	.156	-2.025	.051

	B	Std. Error	T	Sig
常数	.046	.197	.234	.818
接触幸福微笑程度	.039	.314	.124	.903
接触愤怒紧张程度	-.017	.200	-.087	.931
亲密关系接触程度	-.509	.181	-2.807	.013
电视信息接触程度	-.404	.213	-1.892	.077
接触悲伤程度	-.081	.180	-.450	.659

3.2.8.3c 预测睡眠休息程度的多重回归模型

另外，在社会变量中，我们选取了睡眠休息程度作为因变量，力图通过的短信出入的行为变量与通讯情况的社会变量来进行预测。多重回归模型结果建立之后，发现在第一个模型中，5 月份的短信出入情况具有较为有力的解释力度，即收到的短信越少，发出的短信越多，则对应的被试睡眠休息程度越好；而在第二个模型中，与亲戚的电信通讯或见面频率越低，使用网络短信通讯频率越低，则对应的被试睡眠休息程度越好。

表 3.2.8.3c 预测睡眠休息程度的多重回归模型

	B	Std. Error	T	Sig
常数	-.029	.153	-.187	.852
短信in_3	.004	.010	.430	.669
短信out_3	.001	.009	.089	.929
短信in_4	.036	.024	1.472	.148
短信out_4	-.022	.022	-1.016	.315
短信in_5	-.055	.024	-2.282	.027
短信out_5	.040	.021	1.869	.068

	B	Std. Error	T	Sig
常数	.025	.130	.191	.850
使用skype通讯频率	.017	.131	.128	.899
电信通讯或见面频率与亲戚	-.315	.131	-2.411	.020
电信通讯或见面频率与朋友	.166	.132	1.257	.215
使用网络短信通讯频率	-.232	.130	-1.780	.082
使用邮件通讯频率	-.027	.134	-.203	.840

3.3 典型相关分析

经过因子分析后，我们现在所提取的变量详见附录：主要分为大五人格、创造力、通讯情况、社区关系、手机应用程序类别与信息来源的各类评价、共情能力、情绪、生活作息、手机电话、短信、电池信息。通过将不同大类的变量组进行典型相关分析后，我们发现以下结果。

3.3.1 大五人格与社区关系

取第一典型相关变量，典型相关系数为 .519，显著性为 .050。各变量标准典型相关系数如下表所示：

表3.3.1 大五人格与社区关系典型相关分析成分系数

	标准成分系数		标准成分系数
外向性	-0.04	社区融入度	0.072
宜人性	0.031	帮助他人照看孩子频率	-0.293
神经质	-0.003	社区熟识度	0.953
尽责性	1.05		
开放性	-0.215		

我们可以发现尽责性的人格特质和社区熟识度在模型中起到了主要贡献，即相对更加自律、有条理的人会与社区中更多的人进行日常交流，与邻里之间更加熟悉。

3.3.2 大五人格与手机应用程序分类的喜爱或常用程度

取第一典型相关变量，典型相关系数为 .681，显著性为 .047。各变量标准典型相关系数如下表所示：

表3.3.2 大五人格与手机应用程序分类的喜爱或常用程度典型相关分析成分系数

	标准成分系数		标准成分系数
外向性	0.772	生活方式-娱乐	0.097
宜人性	0.439	效率工具-社交平台	0.014
神经质	0.364	游戏-通讯	-0.73
尽责性	0.304	新闻天气	0.677
开放性	0.03		

我们可以发现：在模型中，一边，外向性的人格特质起到了主要贡献；另一边，“新闻天气”类和“游戏减通讯”类（负向）手机应用程序起到了主要贡献。也就是相对更加外向喜欢社交的人会更加喜欢使用新闻天气等程序来了解外部资讯，且更喜欢使用手机上的通讯工具，而不太喜欢宅在家打游戏。

3.3.3 大五人格与创造力

此时，创造力得分为一元变量，典型相关分析相当于做了一次多元回归，模型相关系数为 .836，显著性 < .001。拟合模型如下

$$\text{创造力} = .100 \times \text{外向性} + .218 \times \text{宜人性} - .249 \times \text{神经质} - .122 \times \text{尽责性} + .934 \times \text{开放性}$$

开放性更高的人在创造力表现上也更加出色，说明了创造力和发散性思维存

在较大相关性。

3.3.4 手机应用程序喜爱或常用程度与手机电话、短信数

取第一典型相关变量，典型相关系数为 .745，显著性为 .004。各变量标准典型相关系数如下表所示：

表3.3.4 手机应用程序喜爱或常用程度与手机电话、短信数

	标准成分系数		标准成分系数
生活方式-娱乐	-0.057	接入电话数	-0.293
效率工具-社交平台	-0.418	拨出电话数	1.238
游戏-通讯	0.907	未接电话数	-0.91
新闻天气	0.014	收到短信数	1.274
		发出短信数	-0.484

对比各标准典型相关系数，我们可以推断这里表现出的是平日喜欢打游戏、使用社交平台，而较少使用效率工具和其他通讯程序的手机用户，此类用户呈现出了拨出电话多、未接电话少、收到短信多的趋势。

4.讨论

首先，本次数据研究和分析最主要的意义在于发现了人格五个变量和通讯变量、社区关系、对各类手机应用程序的喜爱与常用程度、对不同来源手机应用程序（低、中、高信任源）的使用频率和价值评价、对自我共情能力的评价、情绪与心境变量、生活方式变量、创造力得分等八组变量的典型相关关系。第一是多元变量间是否存在相关关系，第二是如果有相关关系，这种相关关系可以被哪对或者哪几对变量予以解释。研究表明，人格变量和社区关系、对各类手机应用程序的喜爱与常用程度和创造力得分等三组变量有显著或者边缘显著的典型相关关系。第一，以人格尽责性为主要成分的人格典型变量和以社区熟识度为主要成分的社会关系典型变量的典型相关关系，这个发现说明了尽责性和社区熟识度的相关可以较好地反映人格和社区关系的相关性；第二，以人格的外向性和宜人性为主要成分的人格典型变量和以游戏、新闻手机应用程序喜爱和常用程度为主要成分的对各类手机应用程序的喜爱与常用程度典型变量的典型相关关系；第三，以人格的开放性、外向性、神经质为主要成分的人格典型变量和创造性得分的典型相关系，这其实也是多重回归关系，说明了人格的这三个维度可以较好地预测创造性得分。

典型相关分析的意义在于让我们发现了不同组变量间存在的相关性，也为如何通过人格来预测行为提供了一些建议。比如，并不是人格的所有维度均对所有行为的预测有作用，对于不同的行为和态度变量，需要考虑用不同维度的人格变

量进行预测。

其次，因子分析过程中让我们发现了两点比较有趣的现象。第一，对于 m×n 命名方式的变量来说，因子分析过程可能会出现水平分离现象。例如，在信任源（高，中，低）3×用户体验（使用频率，价值评价）2 的变量中，来自高信任源和低信任源的手机应用程序的使用频率和价值评价被聚到了一起，而来自中信任源的手机应用程序的使用频率和价值评价出现了分离。这样的分离可能是有意义的，且这种意义是因子分析给我们带来的额外收获。以这个变量为例，可能来自中信任源的手机应用程序被试倾向于多使用，但因为信任源受信任度并不高或者品味与被试差异太大，因而价值评价又会偏低。这样的分离给了我们很多信息和进一步研究的猜想与假设。第二，当所聚出来的变量命名上逻辑不统一时，说明其逻辑结构并非平行的。例如，接触幸福微笑程度、接触愤怒紧张程度、亲密关系接触程度、电视信息接触程度这四个因子，前两个是对情绪的共情能力，第三个是对亲密关系的感受能力，最后一个是对媒体信息的理解能力，命名上逻辑并不一致。因而其实可能的逻辑是，对世界的理解和共情能力在本数据集中可以分为对情绪、对人际关系和对媒体信息，其中对情绪又根据效价有显著不同的两方面。因子分析本身对我们了解原变量的信息结构在这一点上起到了很大的帮助。

最后，我们意外地收获了一些没有料想的结果。第一，不同性别的高知识水平研究员在人格上存在着显著的差异，可见知识水平似乎并未对人格能有很好的预测，相反性别角色很大程度决定了人们更可能处于何种人格分布中。第二，聚类分析上，发现夫妻在社区融入度、社区熟识度和帮他人照看孩子等方面能较好地聚为一类。不过我们说不清楚究竟是因为交际程度接近的人容易成为夫妻，还是成为夫妻后，交际程度会变得更为接近。第三，理想的生活方式和实际的生活方式差距显著，深刻印证了“理想很丰满，现实很骨感”这句话。第四，孩子数量可以被多种变量予以预测，尽管目前暂未能由此推出很实用的现实意义，但是这可以是未来研究的一个方向。

我们对数据集做了原有文献所能做到的一些探索，但是我们相信对于这批数据的挖掘还方兴未艾。当加入了新的辅助工具，并对变量有了更好的理解时，GPS、活动范围等数据也能被提炼为具有行为模式意义的变量，进而加入到数据的分析当中去。此外，面对数据集中较多的缺失值，适当地在信息爆炸的当今，改善测量技术，进行重新测量，以获得更新、更完整的数据集，将对电子设备大数据的人格、社会与行为变量相关研究有很大的帮助。

参考文献

[1]Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal and Alex Pentland, "Social fMRI, investigating and shaping social mechanism in the real world", Pervasive and Mobile Computing,2011.

[2]Yves-Alexandre de Montjoye*, Jordi Quoidbach*, Florent Robic*, Alex Sandy Pentland, "Predicting people personality using novel mobile phone-based metrics".Social Computing, Behavioral-Cultural Modeling and Prediction,2013.

[3]Yaniv Altshuler,Michael Fire,Nadav Aharony, Yuval Elovici and Alex Pentland,How Many Makes a Crowd? On the Evolution of Learning as a Factor of Community Coverage, Intl.Conference on Social Computing, Behavioral-Cultural Modeling,and Prediction,2012.

[4]Aniv Altshuler,Michael Fire,Nadav Aharony, Yuval Elovici and Alex Pentland,Social Computing, Behavioral-Cultural Incremental Learning with Accuracy Prediction of Social and Individual Properties from Mobile-Phone Data,Modeling and Prediction,2013.

[5]Sai T Moturu,Inas Khayal, Nadav Aharony, Wei Pan and Alex(Sandy)Pentland Using Social Sensing to Understand the Links Between Sleep, Mood, and Sociability , Socialcom，2011.