Extracting Latent Factors from Survey Data

D.B. Skillicorn
School of Computing

A. Larsen
Department of Sociology

Queen's University

February 2006
External Technical Report
ISSN-0836-0227-
2006-506

School of Computing
Queen's University
Kingston, Ontario, Canada K7L 3N6

Document prepared February 24, 2006
Copyright ©2006 D.. Skillicorn and A. Larsen

Abstract

Surveys requiring Likert scale responses have a number of deficiencies, for example they require a good understanding of possible factors in designing questions, and acceptable answers are often easy to infer. Vignette questions avoid these deficiencies, but can require more sophisticated analysis of responses in order to discover latent or hidden factors which might characterize the space of interest. We describe the use of singular value decomposition as an analysis tool and illustrate the process with a case study of internet use survey data.

Extracting Latent Factors from Survey Data

D.B. Skillicorn and A. Larsen

Abstract: Surveys requiring Likert scale responses have a number of deficiencies, for example they require a good understanding of possible factors in designing questions, and acceptable answers are often easy to infer. Vignette questions avoid these deficiencies, but can require more sophisticated analysis of responses in order to discover latent or hidden factors which might characterize the space of interest. We describe the use of singular value decomposition as an analysis tool and illustrate the process with a case study of internet use survey data.

1 Introduction

Questionnaires are a common tool to discover motivations, attitudes, or beliefs about a particular subject or topic. A conventional questionnaire contains questions that are designed to elicit responses to a set of factors that is assumed to be explain such motivations, attitudes, or beliefs. For example, attitudes to the collection and storage of personal data may be influenced by factors such as individual expectations or requirements for privacy, desire for security (in a government context) or differential service (in a business context), individual models of trust, and individual ideas about what constitutes a community. Typically, questions give the opportunity for respondents to specify how much they agree or disagree with some posited issue related to each factor on a linear scale (say, strongly agree to strongly disagree).

The difficulty with the design of such surveys is that it requires a substantial understanding of the relevant factors, since this is required to generate appropriate questions. In many settings, these factors may be known, and exploring their relationship may be the point of the survey. However, where complex social phenomena are concerned, it may be difficult or impossible to decide, in advance, which factors should be considered. Important factors may only be discovered as a result of considering the totality of survey responses.

Also, since each question addresses a single hypothesised factor, it may be possible for respondents to guess which response is desired, which is socially acceptable, or which best supports their self-image (that is, their rationalized rather than true opinion).

Of course, mechanisms exist to deal with these weaknesses. For example, most factors are not independent, so questions can be constructed that examine aspects or two or more at the same time. Multivariate analysis techniques can assess the extent to which factors are independent or correlated, and assign weightings that describe each factor's overall importance. The tendency of respondents to guess desired answers can be measured, if not discounted, by asking several questions that address the same factors, but in different guises, and comparing the consistency of the responses.

However, the weaknesses of single-factor questions have led to increased interest in the use of questions based on vignettes, short situational stories, to which a set of responses are provided. Vignette questionnaires are especially popular in medical and social work settings [1, 2, 5].

The advantages of vignette questionnaires are:

- Because they are typically in the 3rd person, they remove a reflexive component from the response, perhaps enabling it to be more objective;

- They call for a response to a particular, concrete situation which may evoke a straightforward response, rather than an abstract, hypothetical, or overthought response;
- They can be constructed to be more subtle (that is multifactorial) than simple questionnaires, making it more difficult to guess the desired or socially acceptable response;
- Because they can describe realistic situations with strong actors, they can be more engaging, encouraging full involvement by respondents [6];
- Techniques are known to normalize responses across groups, for example different nationalities [4].

Hence vignette questionnaires can provide better access to the real opinions of respondents, and so may be a better technical tool.

Vignette questionnaires have often been constructed in a similar style to more direct questionnaires, assuming that the underlying factors are already known. For example, each possible response to a vignette is related to a particular factor, and the overall factor weightings are computed from the frequencies of each kind of response.

Stronger forms of factor analysis that are able to extract latent or hidden factors can be applied to both standard Likert-style questionnaires and vignette questionnaires. This stronger analysis means that questionnaires can make much weaker assumptions about the factors that are *supposed* to be present. Instead, vignettes can be constructed to describe situations in which a large number of factors could conceivably be important, and those that are actually important can be discovered, retrospectively, by analysis. These analysis techniques allow available responses to be much more open-ended, since responses do not have to be mapped, *a priori*, to expected factors. They are also able to discard automatically factors that are less significant, so there is little cost to including factors in the vignettes that only *might* be relevant.

In this paper, we present one of these stronger forms of factor analysis, singular value decomposition (SVD), and illustrate its potential in a case study analysing a questionnaire about internet use.

2 Stronger Factor Analysis

The result data from a survey are a set of responses to each vignette by each participant. Suppose that these data are arranged in a table, with one row for each respondent, and one column for each question or vignette. The entries in the table are then the possible responses to each vignette, which might be Likert scale responses, or possible responses to a vignette, which might be selected from a fixed set of responses, but could be, for example, a numerical value indicating strength of response.

For example, consider the following vignette: *Adam tells a friend, Brenda, about some financial irregularity he is involved in at work. Which of the following actions would you consider to have breached Adam's privacy:*

> *Brenda tells a member of Adam's family;*
> *Brenda tells a member of their social circle at a party;*
> *Brenda reveals the information in public by referring to it in a presentation;*
> *Brenda calls Adam's employer and tells them;*
> *Brenda tells the police.*

This vignette raises issues of relationships (social groups), honesty, and privacy in ways that are not easily separable. For people who think that honesty trumps privacy, one of the last two responses is likely. However, even those who think this way might feel themselves bound by the relationship not to go public. Little dependable information about a respondent's beliefs could be learned from this single question, although these could be determined using a series of interlocking vignettes that set up different tensions between these three values.

The responses to these questions could be scored in several ways. For example, the responses could be considered as five distinct (although obviously related) binary responses. They could also be considered as a ranked list (perhaps with the inclusion of a 'None' response) on a scale of 0 to 5, and the largest value recorded as the response. Alternatively, the sum of all of the responses selected could be used as the response value. For the kind of analysis we describe, the particular style of response does not matter, so the analysis can be applied to retrospective survey data without difficulty. However, two issues arise in using existing survey data: first, the direction of a ranked coding must be consistent in order to properly distinguish positive and negative correlation; second, null responses must be assigned values that are meaningful when a response range is used. This latter requirement often requires substantial recoding since it is common to code responses on a scale of 1 to 5 (say) for substantive responses and use 6 to code for 'don't know'. Numerically, this makes the vacuous response seem strong.

In the following, we assume that response data has been tabulated with rows corresponding to individuals and columns corresponding to question or vignette responses. Such a table can be regarded as a matrix, A, with n rows (the respondents) and m columns (the questions). A *singular value decomposition* (SVD) [3] decomposes such a matrix as the product of three matrices, U, S, and V such that

$$A \;=\; U\, S\, V'$$

where U is $n \times m$, S is a diagonal matrix of non-increasing non-negative values, and V is $m \times m$. Here the superscript dash indicates transposition, and both U and V are orthogonal.

The natural interpretation of SVD is geometric. Each row of A can be understood as the coordinates of a point in m-dimensional space. Respondents with similar responses will be close together in this space. However, when m is large, working directly in this space is awkward; for example, relationships cannot be visualized directly. Furthermore, points that are correlated need not be close together (for example, two negatively correlated points will be on opposite sides of the origin). SVD can be thought of as transforming this geometric space into a new one, with a new set of axes (the rows of V') such that as much variation as possible is captured along the first axis, as much as possible of what remains along the second axis, and so on. The coordinates of the point corresponding to a respondent in this new space are given by its row in the U matrix. The responses typically contained redundant information because respondents regard some particular issues in a similar or related way. The new axes can be thought of as describing a set of latent factors, that are as independent as they can be.

However, the biggest benefit of SVD is that it can be truncated by choosing an r smaller than m and discarding all but the first r columns of U, the top left $r \times r$ submatrix of S, and the first r rows of V'. The resulting U matrix can be interpreted as giving the coordinates of points corresponding to each respondent in an r-dimensional space; and this space is the most faithful possible in r dimensions. In particular, if $r = 2$ or 3, visualization of the relative positions of each point become possible. The r axes correspond to the r most important latent factors that underlie the data.

privacy	dataprotect	govtaccess	businessaccess
1	1	4	5
2	1	5	5
4	5	2	2
5	4	2	1
1	5	1	5

Figure 1: An example set of responses by 5 respondents to 4 questions

SVD is completely symmetric with respect to respondents and responses. So it is equally true that the columns of V' correspond to the position of points associated with the m responses in an r-dimensional space. SVD is PCA applied to the respondents and to the responses simultaneously.

There are a number of other ways to interpret SVD. One that is useful in this context is the following: if the points corresponding to the respondents and the points corresponding to the responses are plotted in the same r-dimensional space, then each point lies at the weighted median of the points of the opposite kind (responses for respondents and *vice versa*), where the weights are the entries in the matrix. This justifies the claim that proximity in this r-dimensional space corresponds to affinity between response and respondent, although care must be taken that a large positive response corresponds to *agreement* with the underlying premises of each question.

The axes of the transformed and truncated space, which correspond to the dimensions of this constructed response space, arise automatically from the structure of the data. No prior information about what these dimensions should be is required. Analysis after the fact can often identify an appropriate description for each dimension. These latent factors can either confirm the factors that were believed to be significant, or can suggest factors that better explain the responses that have been collected. The meaning attached to these dimensions also lays the foundation for altering the underlying motivations or beliefs that lead to a response since they are (by construction) orthogonal. This makes it possible to alter responses along one dimension without changing them along other dimensions as a side-effect.

The values on the diagonal of the S matrix arise because there is an inherent ordering of the dimensions – the most significant variation is captured along the first dimension and so on. The magnitudes of these values provide information about the relative importance of these dimensions to the structure of the data. In other words, these values correspond to the importance of each dimension, or how much a global opinion depends on or is consistent with a local opinion with respect to a particular dimension. Also, these values provide the both the motivation and the mechanism for choosing the number of dimensions to retain when truncating – since the magnitudes of these values for the dimensions that are discarded are a measure of how much useful information is being discarded as well.

To illustrate this we look at a small example, the table shown in Figure 1. Suppose the values in this table represent responses to four illustrative questions: do you believe privacy to be important, do you believe businesses should protect your personal data, do you believe governments should be able to access your personal data, and do you believe businesses should be able to access your personal data. High response values indicate strong agreement while low values indicate strong disagreement.

Because of the way the table has been arranged, it is easy to see that the first two respondents give little weight to privacy and are comfortable with both governments and businesses having access to personal data. Respondents 3 and 4 have contrary views; while respondent 5 has views

privacy	govtaccess	businessaccess	dataprotect
1	4	5	1
5	2	1	4
1	1	5	5
4	2	2	5
2	5	5	1

Figure 2: The same data as Figure 1, but without careful ordering; the properties of both respondents and questions are much harder to see

different from both of the other groups and not very internally consistent. Of course, even for such a small table, these relationships would be difficult to see if the rows and columns were rearranged, as shown in Figure 2.

The original table is the A matrix. When we compute the singular value decomposition, we get the following U, Σ and V matrices:

$$U = \begin{bmatrix} -0.51 & -0.04 & -0.58 & 0.45 \\ -0.51 & -0.34 & 0.53 & -0.37 \\ 0.44 & -0.06 & 0.45 & 0.63 \\ 0.53 & -0.41 & -0.42 & -0.43 \\ 0.05 & 0.84 & 0.02 & -0.29 \end{bmatrix}$$

$$S = \begin{bmatrix} 3.35 & 0 & 0 & 0 \\ 0 & 2.13 & 0 & 0 \\ 0 & 0 & 0.44 & 0 \\ 0 & 0 & 0 & 0.15 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.48 & -0.55 & 0.26 & -0.63 \\ 0.52 & 0.45 & 0.68 & 0.28 \\ -0.47 & -0.56 & 0.58 & 0.36 \\ -0.53 & 0.43 & 0.38 & -0.63 \end{bmatrix}$$

Figure 3 shows the relationships among the respondents, correctly showing that respondents 1 and 2, and respondents 3 and 4 hold similar views, while the views of respondent 5 are very different.

Figure 4 shows the relationships among the questions and the issues they address. The most important distinction (along the first dimension labelled V1) is the privacy versus data access dimension. This is the dimension along which there is the most discrimination of the views of the respondents. The next most important distinction (along the second dimension labelled V2) captures the distinction the respondent 5 makes between data protection and business access to data, on the one hand, and privacy and government access to data on the other. This comes out so strongly because the distinctions of the other four respondents are already accounted for in dimension 1.

SVD shows which factors are actually important in survey responses, rather than those that were supposed to be important, and also provides a quantitative assessment of how important each factor is in relation to others. It can be applied to surveys retrospectively to determine whether there are different or more subtle factors in the data. We illustrate this in a case study. It can also

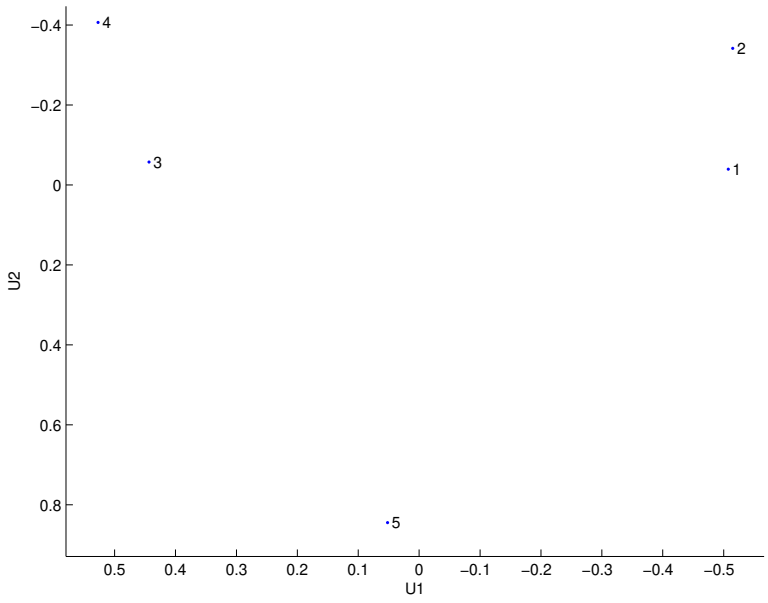

Figure 3: Plot with points corresponding to survey respondents

be applied to results of a prototype questionnaire to determine whether and how the questions may be improved in response to the sample responses.

SVD has the important property that respondents and questions that correlate to no others or that correlate to most others are automatically discounted (placed close to the origin). This means that respondents who respond at random or give stylized responses; and questions that are irrelevant or too bland are automatically discounted in the analysis. This property also gives question designers some flexibility to ask questions in areas where they are unsure of the important factors without contaminating questions in other areas.

Because the factors determined by SVD are orthogonal, they can also be thought of as independent drivers of attitudes or beliefs. In some settings, for example advertising, this provides an opportunity to consider how to change an attitude or belief without unintended consequences.

3 A Case Study

To illustrate this methodology, we consider data from the Georgia Institute of Technology's 10th Annual Worldwide Web User Survey (1998)[1], in which respondents were asked to fill out an online questionnaire describing their internet use, online abilities and experiences, and privacy concerns, using both Likert scale responses and binary yes/no responses. Data from 5022 respondents was used, with 49 responses per respondent (although significant recoding was required to create a table of the appropriate form).

Figure 6 shows a plot of the respondents, using on the first two dimensions, that is coordinates taken from the first two columns of the U matrix of an SVD of the response data. The labels

[1] www.cc.gatech.edu/gvu/user_surveys/

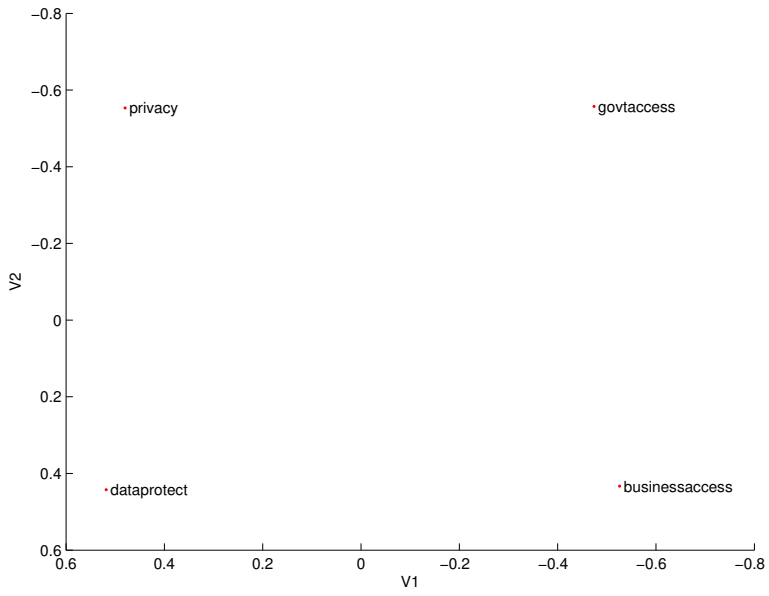

Figure 4: Plot with points corresponding to questions

Figure 5 shows the respondent and response data plotted in the same space, and shows how the points corresponding to respondents are pulled towards questions with which they strongly agree, while the points corresponding to questions are pulled towards respondents who strongly agree with them.

are simply the row number of each respondent, enabling us to relate the position in the plot to a particular set of responses.

The first thing that we observe is that there are separate clusters of respondents, that is there are no abrupt differences among internet users; rather there is a spectrum of similarities for all users. However, the cluster of points is certainly thinner towards the top of the plot.

There are two ways to begin to understand what each dimension in the plot captures. The first is to take individual respondents at the extremes of the cluster, and look at how their responses differ. This can be difficult when there are a large number of responses. However, the results of this analysis can be checked by inserting rows into the table corresponding to artificial respondents whose responses match the putative differences. If the points corresponding to these artificial respondents plot at the expected extremes, then the apparent meaning of the dimension is probably correct.

As mentioned above, in a plot of both respondents and responses, each point is pulled towards points of the opposite kind whenever there is a large data value connecting them. The meaning of the dimensions must therefore agree in both the respondent and response plots. Often, the meaning in the response plot is more obvious.

We can understand the meaning of the first, most significant, dimension, labelled U1 in the plot, by taking points at the left and right extremes and looking at the responses associated with each. For example, the survey asks which of 11 internet-related actions a user has carried out: placed an order online, made a major purchase online, created a web page, customized web page, changed their startup configuration,changed cookie settings,used chat, used internet radio, used

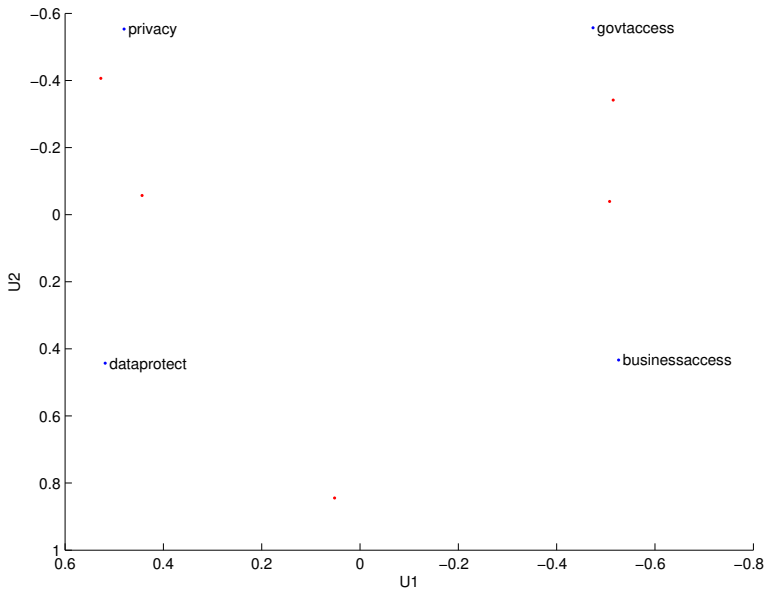

Figure 5: Plot with points corresponding to survey respondents and questions

internet telephone services, used an online directory, attended a seminar, or bought a relevant book. Respondent 101, visible to the right of the plot, has completed 8 of these 11 actions, while respondent 4170, visible to the left of the plot, has carried out only one. Further analysis shows that this is, in fact, the meaning of this first dimension: it represents a continuum from inexperienced (left) to experienced (right). As we might expect, along this dimension there is a continuum of experience, representing steady improvement in users' knowledge of how to carry out various actions on the internet.

We also learn from this plot that experience is *the* most important discriminator among internet users. By considering the values on the diagonal of the S matrix we can quantify this. The first entry is 167, and the second 126, so we can conclude that experience is more important than the next discriminator by 1.3 (=167/126).

We can understand the meaning of the second, vertical, dimension in this Figure in the same way. This dimension corresponds to how internet access is paid for. Respondents in the lower part of the plot either pay for internet access themselves or have it paid for them because of their work, and tend to have access both at home and work. Those in the upper region of the plot have their access paid for by their school or by their parents.

Although it is not surprising that access is one of the discriminators among internet users, it seems surprising that it is such an important one. This finding has, however, been repeated in other internet use datasets that we have analysed. It may be (recall that this is 1998 data) that low-cost access was so limiting that it effectively prevented sophisticated use of the internet or, conversely, that those who wanted to use sophisticated use of the internet were prepared to pay for high-cost access to it.

Figure 7 plots the same data, using the first and third dimensions. The third dimension, labelled by U3, represents the difference in the location of internet access. The lower region of

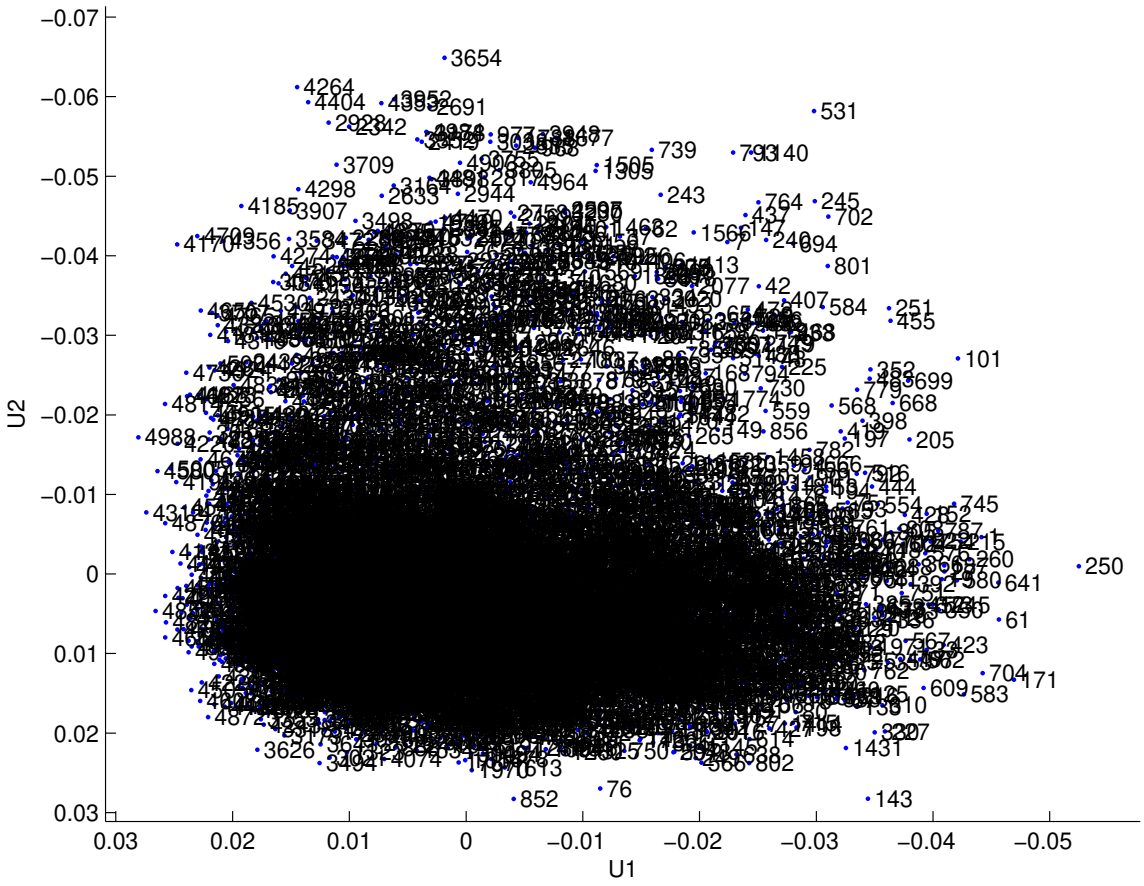

Figure 6: Plot with points corresponding to survey respondents using the two most important underlying factors as dimensions

the plot corresponds to users who access the internet primarily from home, while the upper region corresponds to users who access the internet primarily from work.

Once again, it seems surprising that location of access point should be such an important discriminator among internet users. The third element of the S matrix is 111, so this dimension is 0.88 as important a discriminator as who pays for access. It is also surprising that the two factors of who pays and what kind of access point is used can be separated so well, since it might have been expected that, for example, that access from work would have been strongly correlated with access being paid by work.

Figure 8 provides no new information, but gives a sense of how source of funds and location of internet access interact. In this figure, the bottom left hand corner represents those whose parents or school pay and who access the internet from home; the top left hand corner represents those whose parents or school pay and who access the internet at work (this probably represents primarily students); the top right hand corner represents those whose work pays and who access the internet at work; and the bottom right hand corner represents those whose work pays and

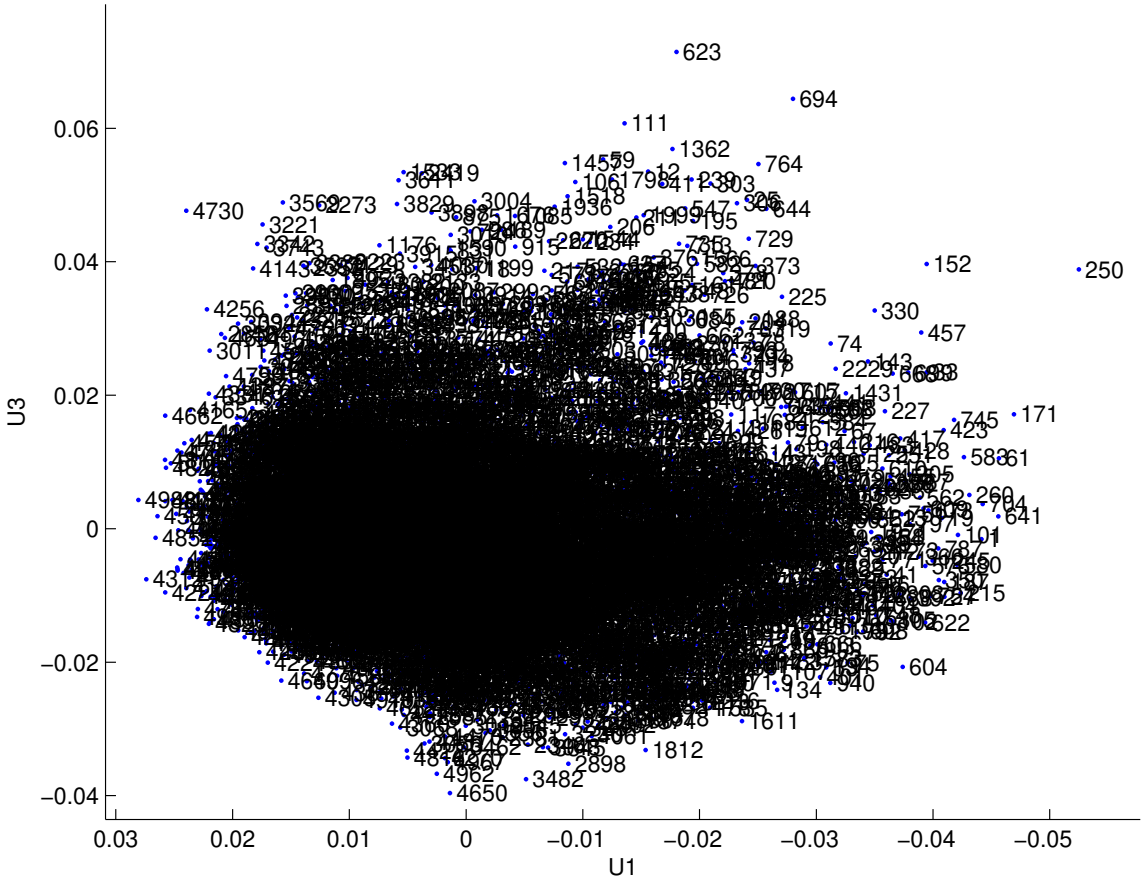

Figure 7: Plot with points corresponding to survey respondents using the first and third most important underlying factors

who access the internet at home. We might call these four regions: children at home, students in residence, employees, and the self-employed, respectively. The dense clump at the right hand end seems to capture individuals who use the internet both at home and at work, and this is consistent with the four regions described above.

We now consider figures that exploit the inherent symmetry of SVD to allow us to plot points corresponding to the questions. These points are labelled as described in Figure 9.

Figure 10 shows a plot of the first two dimensions corresponding to the questions and their responses, with the responses labelled as in Figure 9. The placement of these points reveals much about the similarities among the issues raised by the survey questions. For example, the lower region of the figure shows that greater values for education and income are related, and these in turn are related to paying for one's own internet access. These attributes are also loosely related to age. The left hand side of the figure contains attributes related to internet sophistication and experience: having created or customized a web page, being satisfied with one's internet skills, having used the internet for a number of years and, to a lesser extent, being male. There is an

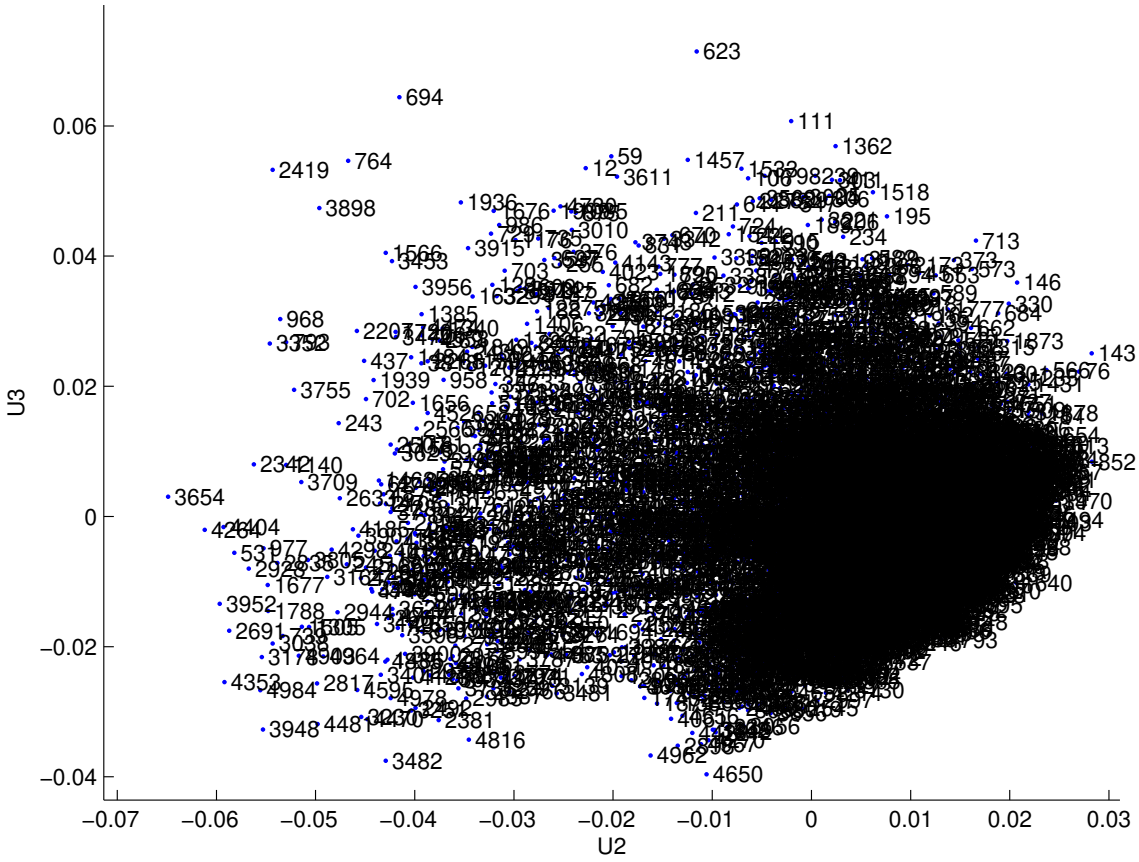

Figure 8: Plot with points corresponding to survey respondents using the second and third most important underlying factors

obvious cluster relating having used the internet to place an order and to make a major purchase, accessing the internet at work, and having the access paid for by work. At the top of the figure, there is another loose cluster of attributes: accessing the internet at school, having the school pay for it, having one's parents pay for it, and having accessed the internet from a public terminal. The right hand side of the figure shows correlation between attributes such as using email, using an unusual platform (the coding for the platform attribute gives high values to WebTV and responses such as 'other' and 'don't know'), being in the non-profit sector or not knowing which sector is appropriate, and being unable to describe one's employer's annual budget. In other words, this region captures attributes that correspond to limited use of the internet within limited contexts.

The structure of the attribute space agrees with that in Figure 6: the first dimension captures experience on the internet, which we are now able to refine into the components that make up 'experience', none of which are very surprising. The second dimension captures who pays for internet access, and we are now able to see that paying oneself and having one's work pay are similar, but that those whose work pays are much more likely to be sophisticated users. On the other hand, those who pay for access themselves are likely to have higher income and be better

educated. Those whose school or parents pay for their access are likely to access the internet from school and to be single, divorced or widowed (because these responses are coded using larger values). In other words, this region probably captures students.

Those attributes whose locations fall close to the origin do not have any significant effect on differentiating internet users. These include whether or not they are registered voters (an odd property in an international context anyway), how many children they have, what kind of industry they work in, where they are located, whether they live in city or country, and what race they are. The irrelevance of some of these is not unexpected, but it seems interesting that so much context has so little effect on internet actions. This suggests that the metaphor of cyberspace as a different kind of universe perhaps has some merit.

Figure 11 is a plot of the question/response attributes in dimensions 1 and 3. This allows us to see which attributes are related to the primary distinction of dimension 3, the difference between access at work and access at home. Access from work and payment for access through work are obviously closely related; they are also correlated with education. The correlation with country seems to arise because values used for coding have small values for U.S. states and Canadian provinces, and then values for countries of the world by region, ending with European countries. Similarly, the coding for language allows explicit answers only for Chinese, Japanese, and European languages, and lumps all other languages in an 'other' category. At the other extreme, there is a correlation between access from home and paying for access oneself, reinforcing the impression that this region represents professionals working from home.

Figure 12 provides a view of the plot where dimensions two and three are simultaneously visible. As before, we can clearly see the close connections between who pays and how access is made with obvious pairs (school pays, access school), (work pays, access work), and (self pays, access home), the last a slightly weaker connection.

Figure 13 plots both the respondents and the questions/respondents in the same space, here using only the first two dimensions. This figure cannot be used directly because there are so many points corresponding to respondents, but it does give some sense of how respondents are defined by their responses – which in the figure means being drawn towards attributes for which their response values were large.

4 Conclusions

There are obvious advantages to vignette-style surveys, since they do not require the underlying factor landscape to be completely understood before survey questions are chosen, and it is harder for respondents to manipulate their responses in response to perceived internal or social pressures. However, the disadvantage of such surveys is that analysis to extract the actual factors and their importance is more difficult. We have suggested singular value decomposition as an appropriate tool, and have illustrated its use by analysing a dataset from an internet use survey. The results show that the most significant discriminator among internet users is their ability and sophistication, which is hardly surprising. However, the next two most important discriminators are who pays for internet access, and where such access takes place. It is much less obvious that these two factors, of all of those available, are important, and this seemed surprising to us. The purpose of the case study, however, was primarily to demonstrate this form of analysis in use.

References

[1] S. Bremberg and T. Nilstun. Patients' autonomy and medical benefit: ethical reasoning among gps. *Family Practice*, 17(2):124–128, April 2000.

[2] S.K. Davidson and P.L. Schattner. Doctor's health-seeking behaviour: A questionnaire survey. *Medical Journal of Australia*, 179:302–305, 2003.

[3] G.H. Golub and C.F. van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

[4] G. King, C.J.L. Murray, J.A. Salomon, and A. Tandon. Enhancing the validity and cross-cultural comparability of survey research. *American Political Science Review*, 97(4), December 2003.

[5] T. Kitamura and F. Kitamura. Reliability of clinical judgment of patients' competency to give informed consent: A case vignette study. *Psychiatry and Clinical Neurosciences*, 54(2):245–247, April 2000.

[6] M. McIntyre. What young women know = what young women do. In *Women's Worlds 99: The 7th International Interdisciplinary Congress on Women*, page 4pp, 1999.

Figure key	Question topic
years	number of years on the internet
lang	primary language
accesshome	frequency of accessing the internet from home
accesswork	frequency of accessing the internet from work
accessschool	frequency of accessing the internet from school
accesspub	frequency of accessing the internet from a public terminal
accessother	frequency of accessing the internet from another place
selfpays	self pays for access
parentspay	parents pay for access
workpays	work pays for access
schoolpays	school pays for access
otherpays	someone else pays for access
dontknowpays	don't know who pays for access
regvoter	registered voter
industry	industry code (arbitrary coding)
occupation	occupation class (arbitrary coding)
sector	occupation sector (arbitrary coding)
emailto	sector where most email goes
budget	size of organization's total budget
education	highest educational attainment
gender	female – male
race	ethnicity (arbitrary coding)
marital	marital status
location	geographical region
residearea	urban–suburban–rural
children	number of children in household
income	household income
age	age
platform	primary computing platform (arbitrary coding)
commbldg	has internet use built a community for you
falsify	how often do you falsify information on the internet
issue	what is the most important issue facing the internet
censorship	do you favour censorship
placedorder	have you placed an order online
majorpurchase	have you made a major purchase online
createdpage	have you created a web page
customizedpage	have you customized a web page
changedstartup	have you changed the startup behaviour
changedcookie	have you changed the cookie behaviour
chat	have you used chat
radio	have you used internet radio
telephone	have you used internet telephone
directory	have you used a directory
seminar	have you attended a seminar
book	have you bought a book
computercomfort	how comfortable are you with computers
internetcomfort	how comfortable are you with the internet
satisfiedskills	how satisfied are you with your skills
country	what state, province, or country do you live in

Figure 9: Meaning of question labels. Survey results were extensively processed so that larger values were associated with increasing sense of the question wherever possible. In general, null or don't know responses were recoded to have minimal impact on the analysis.

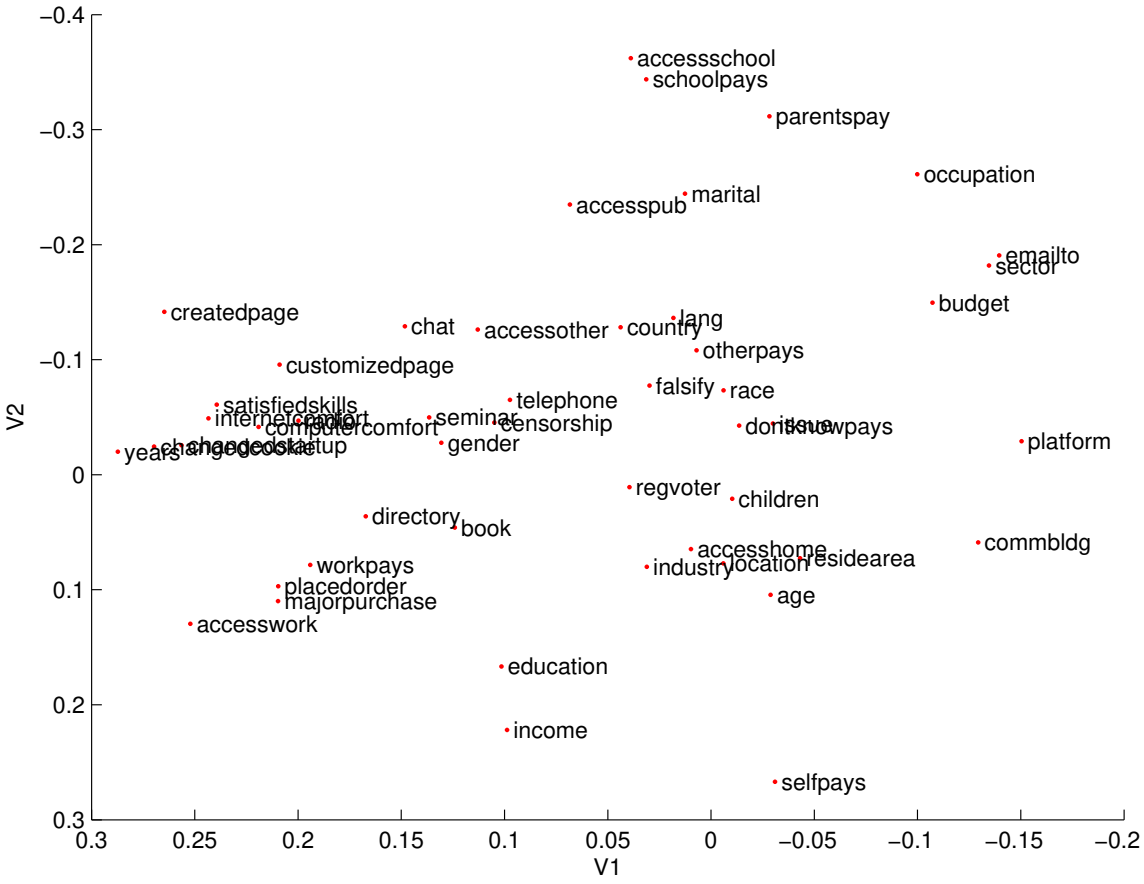

Figure 10: Plot with points corresponding to questions using the two most important underlying factors; labels as in Figure 9

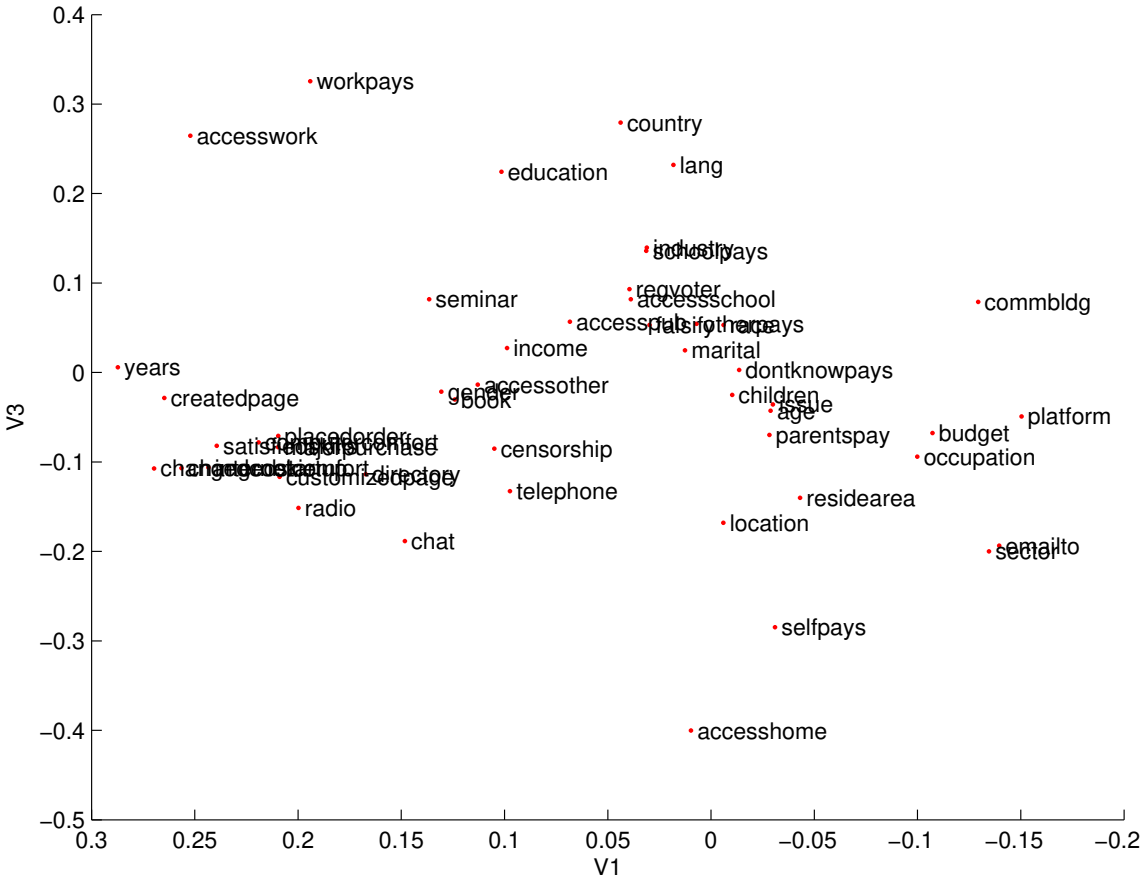

Figure 11: Plot with points corresponding to questions, using the first and third most important underlying factors

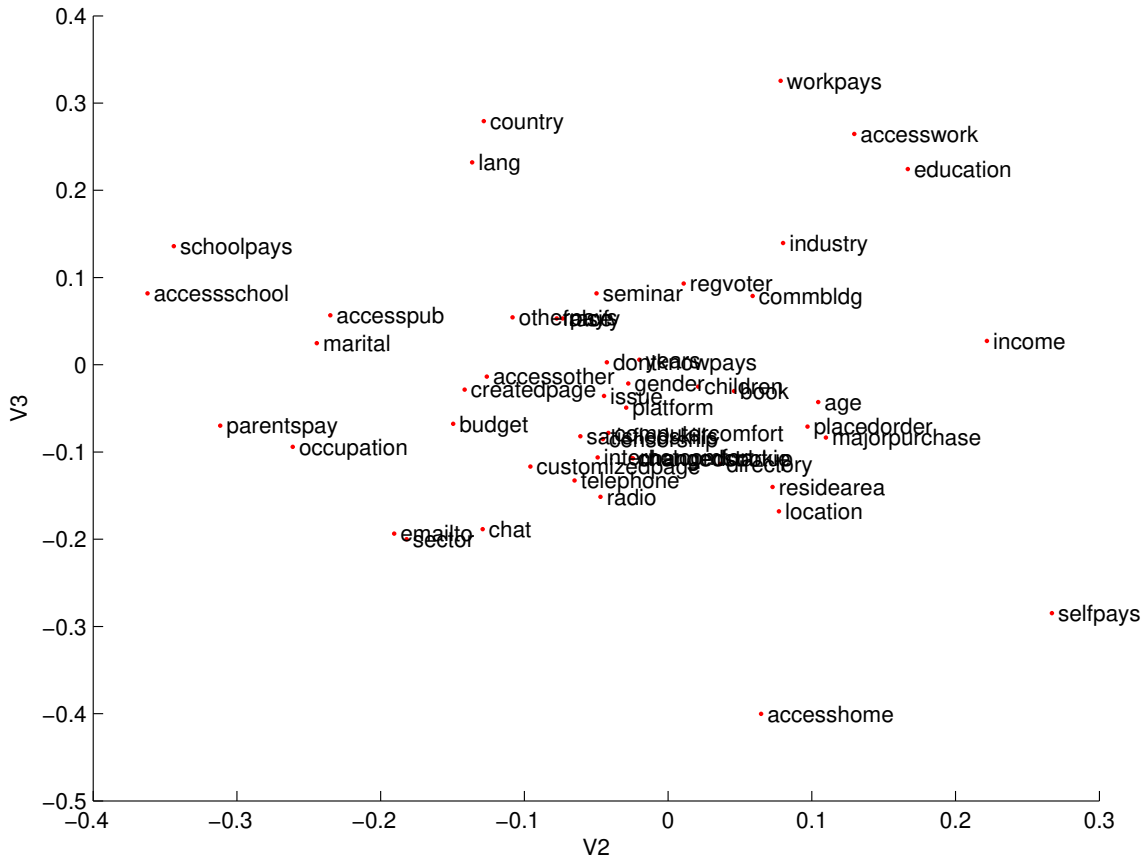

Figure 12: Plot with points corresponding to questions, using the second and third most important underlying factors

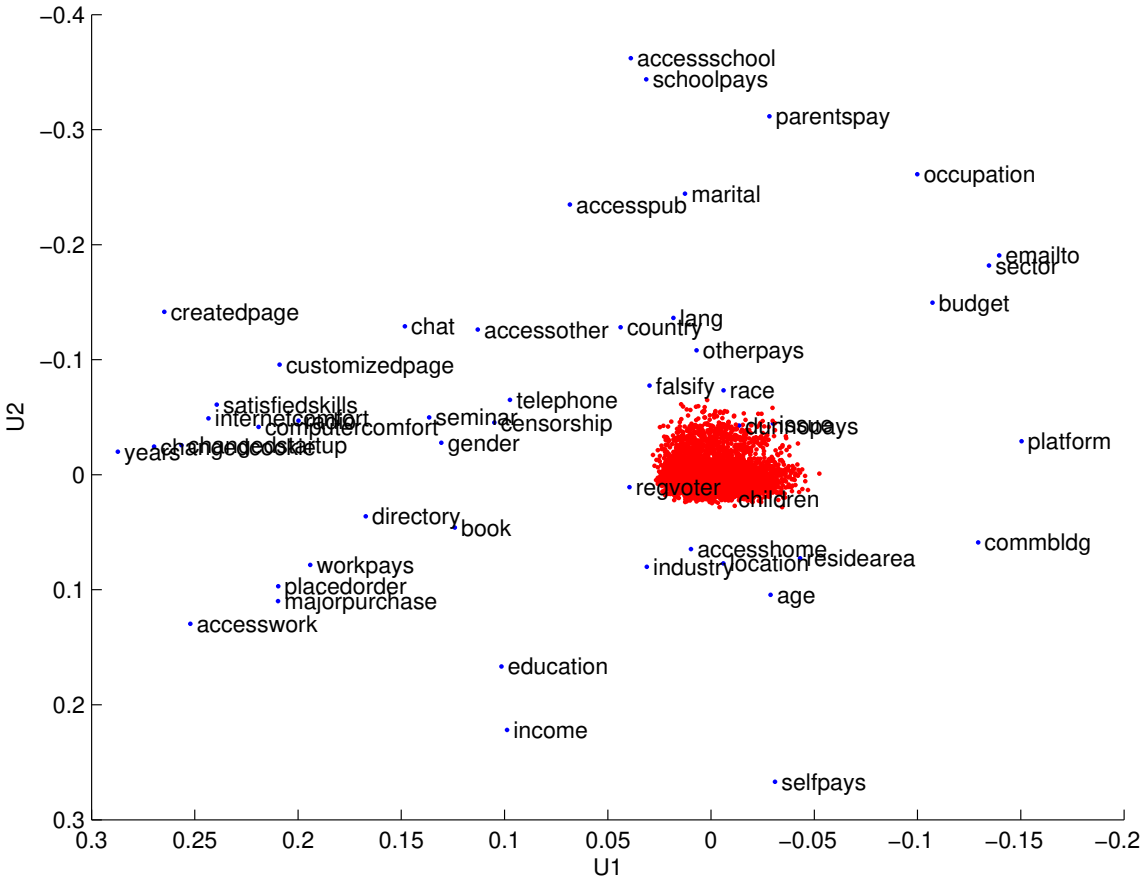

Figure 13: Plot with points corresponding to respondents (red) and questions (blue) using the two most important underlying factors