Psychological Questionnaires and Kernel Extension

Coifman.R.R.⁵  Zucker.S.W.⁶

¹Department of Computer Science, Yale University
²Department of Psychology, University of Haifa.
³Math Department, Davis University.
⁴Engineering School, Bar Ilan University.
⁵Program in Applied Mathematics, Yale University.
⁶Program in Applied Mathematics, Yale University.
Our (online and actual) life, is constantly effected by questionnaires, surveys, tests, etc.

- Personality assessment
- Job Placement
- Psychological evaluation
- Directed marketing
- Online dating and socializing

Can we make sense of people’s answers, without being experts in marketing, dating, or psychology?

In all of the above, We are interested in an underlying property of the responder and NOT in their actual answers.
Answer by YES or NO

**Group A**

- I find it hard to wake up in the morning.
- I’m usually burdened by my tasks for the day.
- I frequently go to wild parties.

And other, that seem less directed:

**Group B?**

- I like poetry.
- I might enjoy being a dog trainer.
- I read the newspaper every day.
These are example questions from The Minnesota Multiphasic Personality Inventory (MMPI-2) which is amongst the most administered psychological evaluation questionnaires in the US.

**Group A** are questions used for estimating depression

- I find it hard to wake up in the morning. (yes)
- I’m usually burdened by my tasks for the day. (yes)
- I frequently go to wild parties. (no)

The depression score is the sum of all indicated matched answers.

**Group B** is designed to test for other conditions and ignored when depression is evaluated.
Our goal:
To learn a scoring function $f$ from answers to scores, using only a training set (answers and scores) and no other prior knowledge.

Advantages:
▶ We are free to learn traits for which we do not know how to manually devise $f$.
  (who is a good employee?)
▶ We are not imposing any ad hoc structure on questionnaire.
  (I might wake up late because I go to wild parties!)
▶ We can limit ourselves to learning noise robust functions.

Drawbacks:
▶ If the property is a complicated (or under determined) function on the answers, we are bound to fail.
The MMPI-2 test

About the MMPI-2:

- It contains 567 questions. (yes/no)
- It evaluates conditions such as: Depression, Hysteria, Paranoia, Schizophrenia, Hypomania, etc.
- Each condition is measured by a scale.
- A scale consists of 10 to 60 questions along with their indicated answers.
- The raw score on a scale is the number of questions answered in the indicated way. (raw scores are normalized to find deviations)
We set:

- Each person’s response to the test is an $x \in \mathbb{R}^{567}$ (yes/no answers $\rightarrow \pm 1$).
- A scoring function $f$, $f_{\text{scale}}(x) : \mathbb{R}^d \rightarrow \mathbb{R}$ is the diagnosis for that person on that scale.

We assume that:

- The scoring function $f$ is sufficiently smooth for a meaningful kernel $K$, id est $\langle f, Kf \rangle \gg 0$.
- The training set sufficiently samples the probability density (and subsequently $f$).
The diffusion kernel is a properly normalized Gaussian kernel. Given a set of \( n \) input vectors \( x_i \in \mathbb{R}^d \)

1. \( K_0(i, j) \leftarrow e^{-\frac{\|x_i-x_j\|^2}{\sigma^2}} \)
2. \( p(i) \leftarrow \sum_{j=1}^{n} K_0(i, j) \) approximates the density at \( x_i \)
3. \( \tilde{K}(i, j) \leftarrow \frac{K_0(i,j)}{p(i)p(j)} \)
4. \( d(i) \leftarrow \sum_{j=1}^{n} \tilde{K}(i, j) \)
5. \( K(i, j) \leftarrow \frac{\tilde{K}(i,j)}{\sqrt{d(i)}\sqrt{d(j)}} \)
6. \( K = USU^T \approx \sum_{k=1}^{m} s_k u_k u_k^T \) (by SVD of \( K \))

Stages 2 and 3 normalize for the density whereas stages 4 and 5 perform the graph laplacian normalization.

Coifman at el. show that in the limit \( n \to \infty, \) and \( \sigma \to 0 \)

- \( K \) converges to a conjugate to the diffusion operator \( \Delta. \)
- The functions \( \varphi_k(x) = u_k(x)/u_1(x) \) converge to the eigenfunctions of \( \Delta. \)
Kernel Extension (Nyström)

Since the $u_k$ are eigenvectors of $K$ we have:

$$\lambda_k u_k(x_i) = \sum_{j=1}^{n} K(x_i, x_j) u_k(x_j) \quad (1)$$

Evaluate $K(x, x_j)$ where $x$ is not in the training set.

$$u_k(x) = \frac{1}{\lambda_k} \sum_{j=1}^{n} K(x, x_j) u_k(x_j) \quad (2)$$

The functions $\varphi_k(x) = u_k(x) / u_1(x)$ extend the kernel to the test set.
Approximating a scoring function $f$

Given a smooth function $f$ over the data points, $f(x_i)$, approximate it with a few $\varphi_k$:

\[ f(x) = \sum_k a_k \varphi_k(x) \]

where
\[ a_k = \int_M \varphi_k(x) f(x) \, dx \approx \sum_{i=1}^n \varphi_k(x_i) f(x_i) p^{-1}(x_i) \, dx \]

$f$ is expressed as a linear combination of $\varphi_k$. We can evaluate $f(x)$ for any $x$.

\[ x \rightarrow K(x, x_i) \rightarrow u_k(k) \rightarrow \varphi_k(x) \rightarrow f(x) \quad (3) \]
Experimental setup and Results

Algorithm parameters:

- $\|x_i - x_j\|$ is the Hamming distance
- Training set size 500 subjects
- Test set size 1000 subjects
- $f$ was approximated by $m = 15$ geometric harmonics

![Figure: Correlations between real and predicted scores for different numbers of geometric harmonics used. For comparison, on the righthand side, the same plot using the PCA kernel eigenfunctions.](image-url)
EASY: Scores over the diffusion map

Depression R.Q = 0.35
Paranoia R.Q = 0.34
Psychasthenia R.Q = 0.31
Hypomania R.Q = 0.35
Psychopathic Deviation R.Q = 0.36
Demoralization R.Q = 0.29

We calculate correlations between the given scores and our predicted scores under three conditions:

1. **EASY**: no missing answers.
2. **HARD**: randomly deleted answers from each test taker.
3. **HARDEST**: delete all answers corresponding to predicted scale. Note, this cannot be scored by other known scoring methods.

When answers are missing the Hamming distance is measured only on the answered questions and scaled up.
It is possible to score accurately with only half the answers!

<table>
<thead>
<tr>
<th>Scale \ missing items</th>
<th>no missing</th>
<th>100</th>
<th>200</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypochondriasis</td>
<td>0.95</td>
<td>0.94</td>
<td>0.93</td>
<td>0.92</td>
</tr>
<tr>
<td>Depression</td>
<td>0.94</td>
<td>0.93</td>
<td>0.93</td>
<td>0.92</td>
</tr>
<tr>
<td>Hysteria</td>
<td>0.89</td>
<td>0.88</td>
<td>0.87</td>
<td>0.85</td>
</tr>
<tr>
<td>Psychopathic Deviation</td>
<td>0.91</td>
<td>0.90</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td>Paranoia</td>
<td>0.87</td>
<td>0.87</td>
<td>0.86</td>
<td>0.84</td>
</tr>
<tr>
<td>Psychasthenia</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Hypomania</td>
<td>0.86</td>
<td>0.86</td>
<td>0.85</td>
<td>0.84</td>
</tr>
<tr>
<td>Social Introversion</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
</tr>
</tbody>
</table>
Scoring Depression with group B equations. All the items belonging to a the predicted scale are missing.

For comparison we tried also to complete the missing responses using a Markov process and score the corrupted records using the usual scoring procedure.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Hit rate</th>
<th>Hit rate MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypochondriasis</td>
<td>59</td>
<td>46</td>
</tr>
<tr>
<td>Depression</td>
<td>67</td>
<td>0</td>
</tr>
<tr>
<td>Hystheria</td>
<td>51</td>
<td>0</td>
</tr>
<tr>
<td>Psychopathic Deviation</td>
<td>59</td>
<td>0</td>
</tr>
<tr>
<td>Paranoia</td>
<td>54</td>
<td>5</td>
</tr>
<tr>
<td>Psychasthenia</td>
<td>88</td>
<td>26</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>85</td>
<td>41</td>
</tr>
<tr>
<td>Hypomania</td>
<td>58</td>
<td>2</td>
</tr>
<tr>
<td>Social Introversion</td>
<td>69</td>
<td>7</td>
</tr>
</tbody>
</table>

Table: Correlations and hit rates variance, for different choices of a training set, is smaller the 0.02.
The same algorithm was run on another MMPI-2 data set, and on dating service data, with similar results.

Psychological questionnaires and their scoring can be addressed with kernel extension ideas.

Filling in missing data might be the wrong thing to do.
Thank you.
Table: $r$, Correlation between real and predicted score. $q$, number of randomly deleted items. The hit rate indicated is the percent of subjects classified within $1/2$ standard deviation from their original score. Correlations and hit rates variance, for different choices of a training set, is smaller the 0.02.