# Psychological Questionnaires and Kernel Extension

Liberty.E[1] Almagor.M [2] James.B [3] Keller.Y [4]
Coifman.R.R. [5] Zucker.S.W. [6]

[1]Department of Computer Science, Yale University
[2]Department of Psychology, University of Haifa.
[3]Math Department, Davis University.
[4]Engineering School, Bar Ilan University.
[5]Program in Applied Mathematics, Yale University.
[6]Program in Applied Mathematics, Yale University.

# Psychological Questionnaires

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, with out 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.**

# Psychological Questionnaires

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.

# Psychological Questionnaires

These are example questions from The Minnesota Multiphasic Personality Inventory (MMPI-2) which is amongst the most administered psychological evaluation questionnaires is the US.

**Group A** are questions are 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.

# Approach, Advantages, and Drawbacks

**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)

# MMPI-2 and the kernel framework

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 $f$, $f_{scale}(x) : \mathbb{R}^d \rightarrow \mathbb{R}$ 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 $f$).

# Diffusion kernel and eigenfunctions

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. $\widetilde{K}(i,j) \leftarrow \frac{K_0(i,j)}{p(i)p(j)}$
4. $d(i) \leftarrow \sum_{j=1}^{n} \widetilde{K}(i,j)$
5. $K(i,j) \leftarrow \frac{\widetilde{K}(i,j)}{\sqrt{d(i)}\sqrt{d(j)}}$
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 $n \rightarrow \infty$, and $\sigma \rightarrow 0$

- $K$ 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:

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

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

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

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 $\varphi_k$:

$$\begin{aligned} f(x) &= \sum_k a_k \varphi_k(x) \\ \text{where} \quad a_k &= \int_M \varphi_k(x) f(x) dx \\ &\approx \sum_{i=1}^{n} \varphi_k(x_i) f(x_i) p^{-1}(x_i) dx \end{aligned}$$

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

$$x \rightarrow K(x, x_i) \rightarrow u_k(k) \rightarrow \varphi_k(x) \rightarrow f(x) \tag{3}$$

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

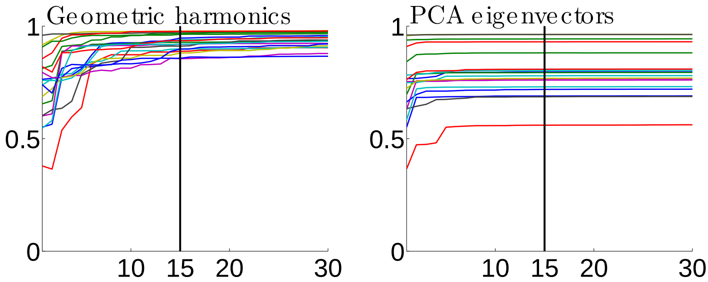
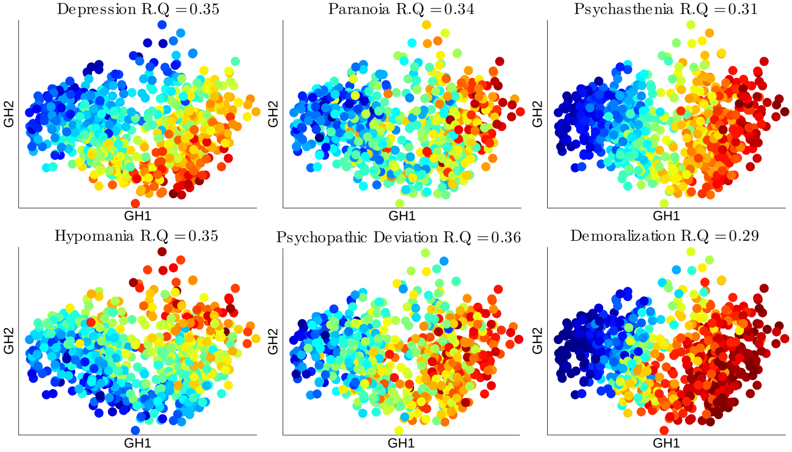



Figure: Correlations between real and predicted scores for different numbers of geometric harmonics used. For comparison, on the righthand side, the same plot using the PCA kernel eigenfunctions.

# EASY: Scores over the diffusion map



# Missing data

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.

# EASY AND HARD: Data missing at random

It is possible to score accurately with only half the answers!

| Scale \ missing items | no missing | 100 | 200 | 300 |
|---|---|---|---|---|
| Hypochondriasis | 0.95 | 0.94 | 0.93 | 0.92 |
| Depression | 0.94 | 0.93 | 0.93 | 0.92 |
| Hysteria | 0.89 | 0.88 | 0.87 | 0.85 |
| Psychopathic Deviation | 0.91 | 0.90 | 0.90 | 0.88 |
| Paranoia | 0.87 | 0.87 | 0.86 | 0.84 |
| Psychasthenia | 0.98 | 0.98 | 0.97 | 0.97 |
| Schizophrenia | 0.98 | 0.98 | 0.97 | 0.97 |
| Hypomania | 0.86 | 0.86 | 0.85 | 0.84 |
| Social Introversion | 0.97 | 0.96 | 0.96 | 0.95 |

# HARDEST: Missing entire scale

Scoring Depression with **group B** equations.
All the items belonging to a the predicted scale are missing.

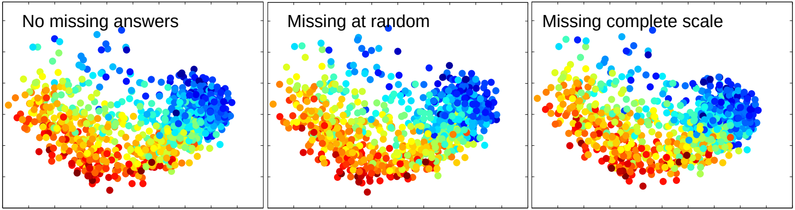
For comparison we tried also to complete the missing responses using a Markov process and score the corrupted records using the usual scoring procedure.

| Scale | $r$ | Hit rate | $r_{MC}$ | Hit rate$_{MC}$ |
|---|---|---|---|---|
| Hypochondriasis | 79 | 59 | 69 | 46 |
| Depression | 86 | 67 | 65 | 0 |
| Hysteria | 74 | 51 | 55 | 0 |
| Psychopathic Deviation | 80 | 59 | 48 | 0 |
| Paranoia | 78 | 54 | 55 | 5 |
| Psychasthenia | 94 | 88 | 70 | 26 |
| Schizophrenia | 94 | 85 | 73 | 41 |
| Hypomania | 80 | 58 | 35 | 2 |
| Social Introversion | 87 | 69 | 58 | 7 |

Table: Correlations and hit rates variance, for different choices of a training set, is smaller the 0.02.

# Summary points

Figure: Depression on the diffusion map. EASY → HARD → HARDEST




- 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.

# Answers missing at random

| | q=30 | | q=50 | | q=100 | | q=200 | | q=300 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $r$ | Hit rate | $r$ | Hit rate | $r$ | Hit rate | $r$ | Hit rate | $r$ | Hit rate |
| HS | 95 | 89 | 95 | 89 | 94 | 87 | 94 | 83 | 92 | 80 |
| D | 93 | 83 | 93 | 83 | 93 | 83 | 92 | 80 | 92 | 77 |
| HY | 89 | 71 | 88 | 70 | 88 | 69 | 87 | 67 | 84 | 62 |
| PD | 91 | 76 | 91 | 77 | 91 | 76 | 90 | 74 | 89 | 71 |
| PA | 88 | 67 | 88 | 67 | 87 | 66 | 87 | 64 | 85 | 62 |
| PT | 98 | 97 | 98 | 97 | 98 | 97 | 98 | 97 | 97 | 96 |
| SC | 98 | 98 | 98 | 98 | 98 | 98 | 98 | 98 | 97 | 97 |
| MA | 87 | 67 | 87 | 67 | 86 | 67 | 86 | 64 | 85 | 65 |
| SI | 96 | 91 | 96 | 92 | 96 | 91 | 95 | 90 | 95 | 88 |
| RCD | 98 | 96 | 97 | 96 | 97 | 96 | 97 | 94 | 97 | 94 |
| RC1 | 95 | 86 | 94 | 86 | 94 | 85 | 93 | 81 | 91 | 75 |
| RC2 | 93 | 82 | 93 | 82 | 92 | 80 | 92 | 79 | 91 | 77 |
| RC3 | 89 | 73 | 89 | 73 | 89 | 72 | 89 | 71 | 88 | 70 |
| RC4 | 92 | 81 | 92 | 78 | 92 | 76 | 90 | 74 | 88 | 69 |
| RC6 | 92 | 78 | 92 | 78 | 91 | 78 | 91 | 75 | 89 | 72 |
| RC7 | 96 | 92 | 96 | 93 | 96 | 92 | 96 | 91 | 95 | 91 |
| RC8 | 93 | 84 | 93 | 84 | 93 | 83 | 93 | 82 | 91 | 78 |
| RC9 | 93 | 82 | 93 | 81 | 92 | 80 | 92 | 77 | 91 | 75 |

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