Scoring Psychological Questionnaires using Geometric Harmonics

Liberty.E[♮], Almagor.M[†], Keller.Y[‡], Coifman.R.R.[§], Zucker.S.W.[♯]

[♮] Department of Computer Science, Yale University.
[†] Department of Psychology, University of Haifa.
[‡] Electrical & Computer Engineering Department, Ben-Gurion University.
[§] Program in Applied Mathematics, Yale University.
[♯] Program in Applied Mathematics, Yale University.

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 love dancing.

What about **Group B**?

- I like poetry.
- I might enjoy being a dog trainer.
- I read the newspaper every day.

Psychological Questionnaires

Group A are questions like the ones in the MMPI-2 test, aimed at 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 love dancing. (no)

In the MMPI-2 a (raw) score is the sum of "correct answers".

Group B, designed to test for other conditions, seem unrelated to depression.

- I like poetry. (?)
- I might enjoy being a dog trainer. (?)
- I read the newspaper every day. (?)

Psychological Questionnaires

Questions:

- Are **Group B** answers informative about depression?
- If so, can incomplete questionnaires be scored correctly?
- Is the space of answers structured? and How?

Answering the latter suggests an approach to the former.

- MMPI-2 structure
- Manifold learning

MMPI-2 and the diffusion framework

- Ambient space: $x \in 567$ dimensions (yes/no answers $\rightarrow \pm 1$).
- A set of responses x_i lie on or near a low dimensional manifold M in $\mathbb{R}^d$
- M is sufficiently sampled with some density p by the training set. For a given function g and a compact subset of $\mathbb{R}^d$,Ω:

$$\sum_{x_i \in \Omega} g(x_i) \approx \int_{\Omega \cap M} g(x) p(x) d\Omega \tag{1}$$

MMPI-2 and the diffusion framework

- scales: functions on the answer vectors $f_{diagnosis}(x) : \mathbb{R}^d \to \mathbb{R}$. summation of "correct answers".
- diagnosis $\in \{$ anxiety, depression, $\ldots$, hysteria $\}$.
- The scoring function $f : \mathbb{R}^d \to \mathbb{R}$ is smooth on M.

$$\text{Raleigh Quotient (R.Q).} \quad \frac{\langle f, \Delta f \rangle}{\langle f, f \rangle} \gg 0 \tag{2}$$

Gaussian kernel normalization

The assumption that high dimensional data reside on or near a low dimensional manifold inspired many theoretical and experimental results.

Short history of the gaussian kernel in dimensionality reduction:

- Schölkopf and Samola used uses the gaussian kernel with no normalization for non-linear PCA .
- Belkin and Niyogi normalize the gaussian kernel to be the laplacian of a graph defined on the data.
- Coifman and Lafon further normalize for non-uniform sampling from the manifold.

Diffusion kernel and eigenfunctions

Given a set of n input vectors $x_i \in \mathbb{R}^d$

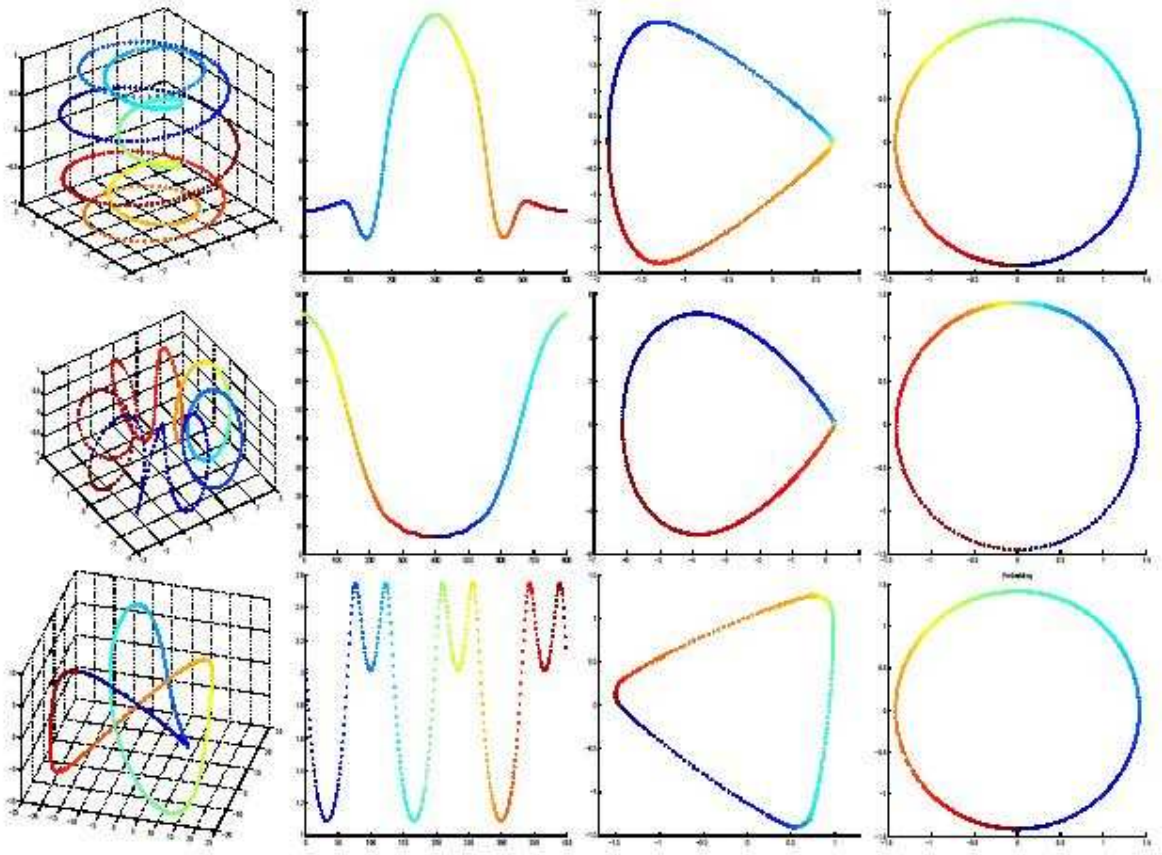
1. $K_0(i,j) \leftarrow e^{-\frac{\|x_i - x_j\|^2}{\sigma^2}}$
2. $p(i) \leftarrow \sum_{j=1}^{n} K_0(i,j)$ approximates the density at x_i
3. $\widetilde{K}(i,j) \leftarrow \frac{K_0(i,j)}{p(i)p(j)}$
4. $d(i) \leftarrow \sum_{j=1}^{n} \widetilde{K}(i,j)$
5. $K(i,j) \leftarrow \frac{\widetilde{K}(i,j)}{\sqrt{d(i)}\sqrt{d(j)}}$
6. $USU^T = K$ (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 Δ.
- The functions $\varphi_k(x) = u_k(x)/u_0(x)$ converge to the eigenfunctions of Δ on M.

Importance of normalization



1D manifold in 3D	Density vs. arc length	laplacian normalization	diffusion normalization

Geometric harmonics, Nyström extension

Since the u_k are eigenvectors of K we have:

$$\lambda_k u_k(x_i) = \sum_{j=1}^{n} K(x_i, x_j) u_k(x_j) \tag{3}$$

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

$$u_k(x) = \frac{1}{\lambda_k} \sum_{j=1}^{n} K(x, x_j) u_k(x_j) \tag{4}$$

The functions $\varphi_k(x) = u_k(x)/u_0(x)$ are therefore evaluated for any x near the manifold. $\varphi_k(x)$ are termed *geometric harmonics*.

Approximating a scoring function f

Given a smooth function f over the data points, $f(x_i)$, approximate it with a few φ_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 now expressed as a linear combination of φ_k. Since we already know how to evaluate $\varphi_k(x)$ for all k and x we can evaluate $f(x)$ for any x.

Summary of geometric harmonics

- We saw how to approximate the eigenfunctions of Δ under non uniform sampling of the training set.
- We defined the *geometric harmonics* and saw how to evaluate them on points outside the training set.
- We saw how to express a function on the training data f as a linear combination of geometric harmonics.
- Finally we combined steps 2 and 3 to evaluate f on new points outside the training set.

Experimental setup for the MMPI-2

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

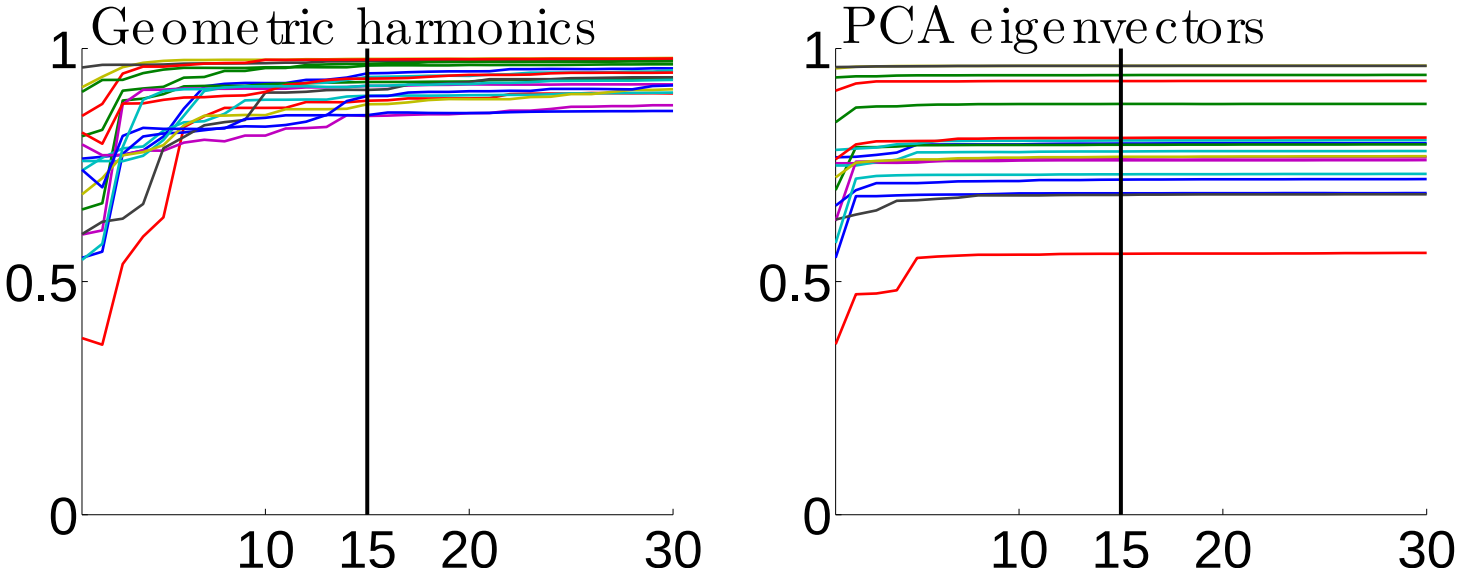
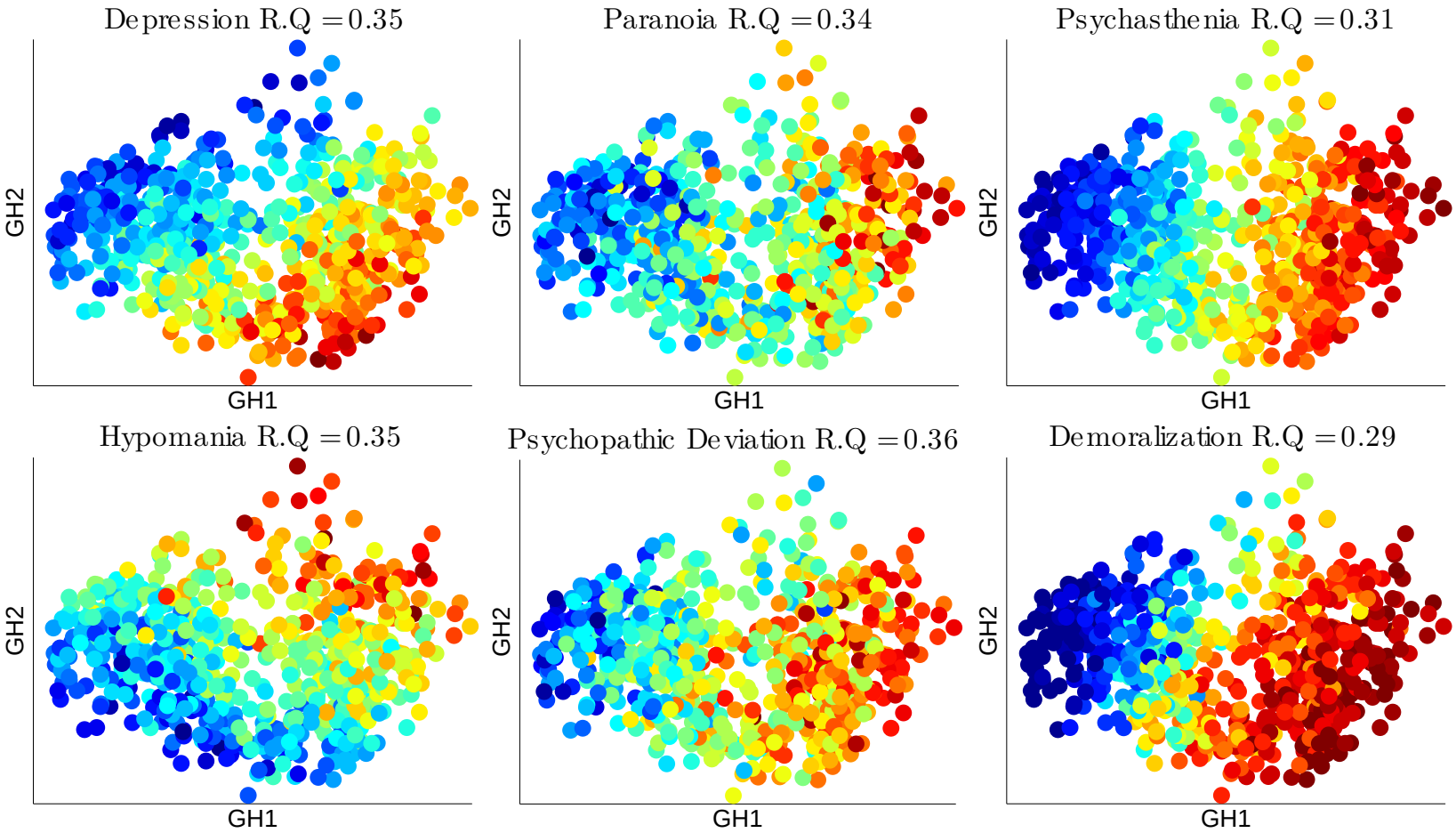


Figure: Correlations between real and predicted scores for different numbers of geometric harmonics used. For comparison, on the righthand side, the same plot using the PCA kernel eigenfunctions.

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.

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.

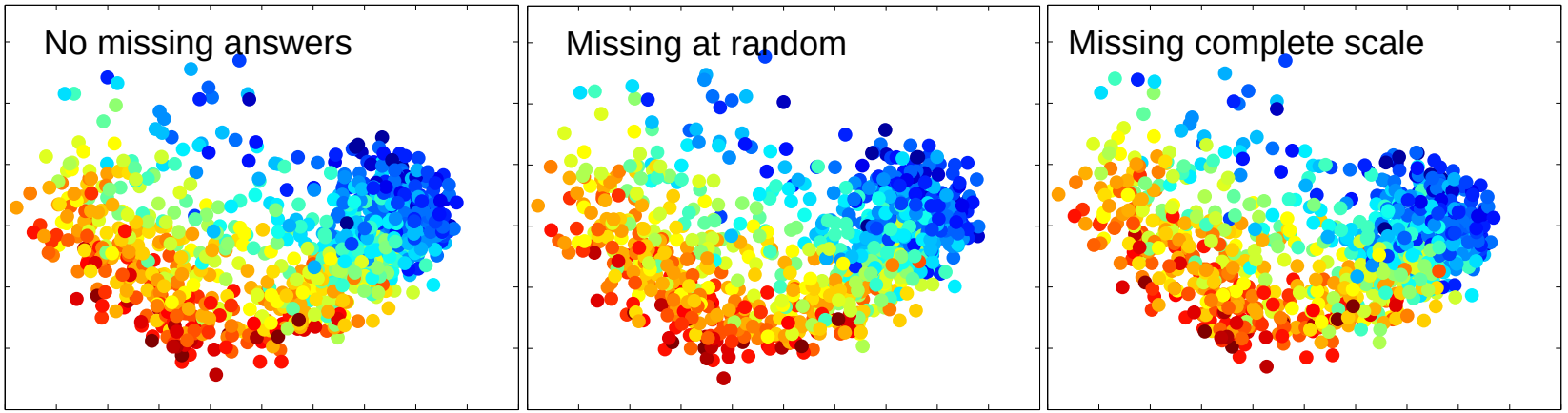
For comparison we tried also to complete the missing responses using a Markov process and score the corrupted records using the usual scoring procedure.

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

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



- Geometric harmonics proved to be a useful tool in scoring personality questionnaires.
- This method is superior to others when dealing with corrupted responses.
- Discrete psychological questionnaires and their scoring fall under the manifold learning framework.

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

Scale definitions

1. Scale 1 Hypochondriasis Neurotic concern over bodily functioning.
2. Scale 2 Depression Poor morale, lack of hope in the future, and a general dissatisfaction with one's own life situation. High scores are clinical depression whilst lower scores are more general unhappiness with life.
3. Scale 3 Hysteria Hysterical reaction to stressful situations. Often have 'normal' facade and then go to pieces when faced with a 'trigger' level of stress. People who tend to score higher include brighter, better educated and from higher social classes. Women score higher too.
4. Scale 4 Psychopathic Deviation Measures social deviation, lack of acceptance of authority, amorality. Adolescents tend to score higher.
5. Scale 5 Masculinity-Femininity This scale was originally developed to identify homosexuals, but did not do so accurately. Instead, it is used to measure how strongly an individual identifies with the traditional (pre-1960's) masculine or feminine role. Men tend to get higher scores. It is also related to intelligence, education, and socioeconomic status.
6. Scale 6 Paranoia Paranoid symptoms such as ideas of reference, feelings of persecution, grandiose self-concepts, suspiciousness, excessive sensitivity, and rigid opinions and attitudes.
7. Scale 7 Psychasthenia Originally characterized by excessive doubts, compulsions, obsessions, and unreasonable fears, it now indicates conditions such as Obsessive Compulsive Disorder (OCD). It also shows abnormal fears, self-criticism, difficulties in concentration, and guilt feelings.
8. Scale 8 Schizophrenia Assesses a wide variety of content areas, including bizarre thought processes and peculiar perceptions, social alienation, poor familial relationships, difficulties in concentration and impulse control, lack of deep interests, disturbing questions of self-worth and self-identity, and sexual difficulties.
9. Scale 9 Hypomania Tests for elevated mood, accelerated speech and motor activity, irritability, flight of ideas, and brief periods of depression.
10. Scale 0 Social Introversion Tests for a person's tendency to withdraw from social contacts and responsibilities.