Algorithms and Data Structures

**1.Factor Analysis**

1. The reason for choose Factor Analysis

Factor analysis is a useful tool for investigating variable relationships for complex concepts such as socioeconomic status, dietary patterns, or psychological scales. It allows researchers to investigate concepts that are not easily measured directly by collapsing a large number of variables into a few interpretable underlying factors.

What has contributed to this continued increase in the use of factor analysis in the health sciences in particular? Several possibilities come to mind:

* Increased researcher interest in the complex organizational structure of various health-related constructs
* Recent developments in the use of confirmatory factor analysis and structural equation modeling
* Greater sophistication concerning statistics on the part of some health care researchers from all disciplines and levels of expertise
* Increased availability of inexpensive but powerful personal computers, which can undertake analyses quickly and inexpensively
* Availability of increasingly user-friendly statistical computer packages

1. Mathematical model

In order for the variables to be on equal footing, they are normalized:

z a i = x a i − μ a σ a {\displaystyle z\_{ai}={\frac {x\_{ai}-\mu \_{a}}{\sigma \_{a}}}}

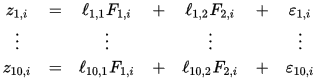
where the sample mean is:

μ a = 1 N i ∑ i x a i {\displaystyle \mu \_{a}={\tfrac {1}{N\_{i}}}\sum \_{i}x\_{ai}} 

and the sample variance is given by:

σ a 2 = 1 N i ∑ i ( x a i − μ a ) 2 {\displaystyle \sigma \_{a}^{2}={\tfrac {1}{N\_{i}}}\sum \_{i}(x\_{ai}-\mu \_{a})^{2}} 

The factor analysis model for this particular sample is then:

z 1 , i = ℓ 1 , 1 F 1 , i + ℓ 1 , 2 F 2 , i + ε 1 , i ⋮ ⋮ ⋮ ⋮ z 10 , i = ℓ 10 , 1 F 1 , i + ℓ 10 , 2 F 2 , i + ε 10 , i {\displaystyle {\begin{matrix}z\_{1,i}&=&\ell \_{1,1}F\_{1,i}&+&\ell \_{1,2}F\_{2,i}&+&\varepsilon \_{1,i}\\\vdots &&\vdots &&\vdots &&\vdots \\z\_{10,i}&=&\ell \_{10,1}F\_{1,i}&+&\ell \_{10,2}F\_{2,i}&+&\varepsilon \_{10,i}\end{matrix}}} 

or, more succinctly:

z a i = ∑ p ℓ a p F p i + ε a i {\displaystyle z\_{ai}=\sum \_{p}\ell \_{ap}F\_{pi}+\varepsilon \_{ai}} 

where

* F 1 , i {\displaystyle F\_{1,i}}  is the ith student's "verbal intelligence",
* F 2 , i {\displaystyle F\_{2,i}}  is the ith student's "mathematical intelligence",
* ℓ a p {\displaystyle \ell \_{ap}}  are the factor loadings for the ath subject, for p = 1, 2.

In matrix notation, we have

Z = L F + ϵ {\displaystyle Z=LF+\epsilon } 

Observe that by doubling the scale on which "verbal intelligence"—the first component in each column of F—is measured, and simultaneously halving the factor loadings for verbal intelligence makes no difference to the model. Thus, no generality is lost by assuming that the standard deviation of verbal intelligence is 1. Likewise for mathematical intelligence. Moreover, for similar reasons, no generality is lost by assuming the two factors are uncorrelated with each other. In other words:

∑ i F p i F q i = δ p q {\displaystyle \sum \_{i}F\_{pi}F\_{qi}=\delta \_{pq}} 

where δ p q {\displaystyle \delta \_{pq}}  is the [Kronecker delta](https://en.wikipedia.org/wiki/Kronecker_delta) (0 when p ≠ q {\displaystyle p\neq q}  and 1 when p = q {\displaystyle p=q} ).The errors are assumed to be independent of the factors:

∑ i F p i ε a i = 0 {\displaystyle \sum \_{i}F\_{pi}\varepsilon \_{ai}=0} 

Note that, since any rotation of a solution is also a solution, this makes interpreting the factors difficult. See disadvantages below. In this particular example, if we do not know beforehand that the two types of intelligence are uncorrelated, then we cannot interpret the two factors as the two different types of intelligence. Even if they are uncorrelated, we cannot tell which factor corresponds to verbal intelligence and which corresponds to mathematical intelligence without an outside argument.

The values of the loadings L, the averages μ, and the [variances](https://en.wikipedia.org/wiki/Variance) of the "errors" ε must be estimated given the observed data X and F (the assumption about the levels of the factors is fixed for a given F). The "fundamental theorem" may be derived from the above conditions:

∑ i z a i z b i = ∑ p ℓ a p ℓ b p + ∑ i ε a i ε b i {\displaystyle \sum \_{i}z\_{ai}z\_{bi}=\sum \_{p}\ell \_{ap}\ell \_{bp}+\sum \_{i}\varepsilon \_{ai}\varepsilon \_{bi}} 

The term on the left is just the correlation matrix of the observed data, and its  diagonal elements will be 1's. The last term on the right will be a diagonal matrix with terms less than unity. The first term on the right is the "reduced correlation matrix" and will be equal to the correlation matrix except for its diagonal values which will be less than unity. These diagonal elements of the reduced correlation matrix are called "communalities":

h a 2 = 1 − ψ a = ∑ p ℓ a p ℓ a p {\displaystyle h\_{a}^{2}=1-\psi \_{a}=\sum \_{p}\ell \_{ap}\ell \_{ap}} 

The sample data z a i {\displaystyle z\_{ai}}  will not, of course, exactly obey the fundamental equation given above due to sampling errors, inadequacy of the model, etc. The goal of any analysis of the above model is to find the factors F p i {\displaystyle F\_{pi}}  and loadings ℓ a p {\displaystyle \ell \_{ap}}  which, in some sense, give a "best fit" to the data. In factor analysis, the best fit is defined as the minimum of the mean square error in the off-diagonal residuals of the correlation matrix:

ε 2 = ∑ a b , a ≠ b [ ∑ i z a i z b i − ∑ p ℓ a p ℓ b p ] 2 {\displaystyle \varepsilon ^{2}=\sum \_{ab,a\neq b}\left[\sum \_{i}z\_{ai}z\_{bi}-\sum \_{p}\ell \_{ap}\ell \_{bp}\right]^{2}} 

This is equivalent to minimizing the off-diagonal components of the error covariance which, in the model equations have expected values of zero. This is to be contrasted with principal component analysis which seeks to minimize the mean square error of all residuals. Before the advent of high speed computers, considerable effort was devoted to finding approximate solutions to the problem, particularly in estimating the communalities by other means, which then simplifies the problem considerably by yielding a known reduced correlation matrix. This was then used to estimate the factors and the loadings. With the advent of high-speed computers, the minimization problem can be solved quickly and directly, and the communalities are calculated in the process, rather than being needed beforehand. The [MinRes](https://en.wikipedia.org/wiki/Generalized_minimal_residual_method) algorithm is particularly suited to this problem, but is hardly the only means of finding an exact solution.

Quoted from Wikipedia: <https://en.wikipedia.org/wiki/Factor_analysis>

**2.Data Structure**

Our data are collected from Fitbit interface and healthy data website. Moreover, we must collect data from personal social networks. All these data will be stored in MySQL database. Here are some data diagrams to describe data structure:

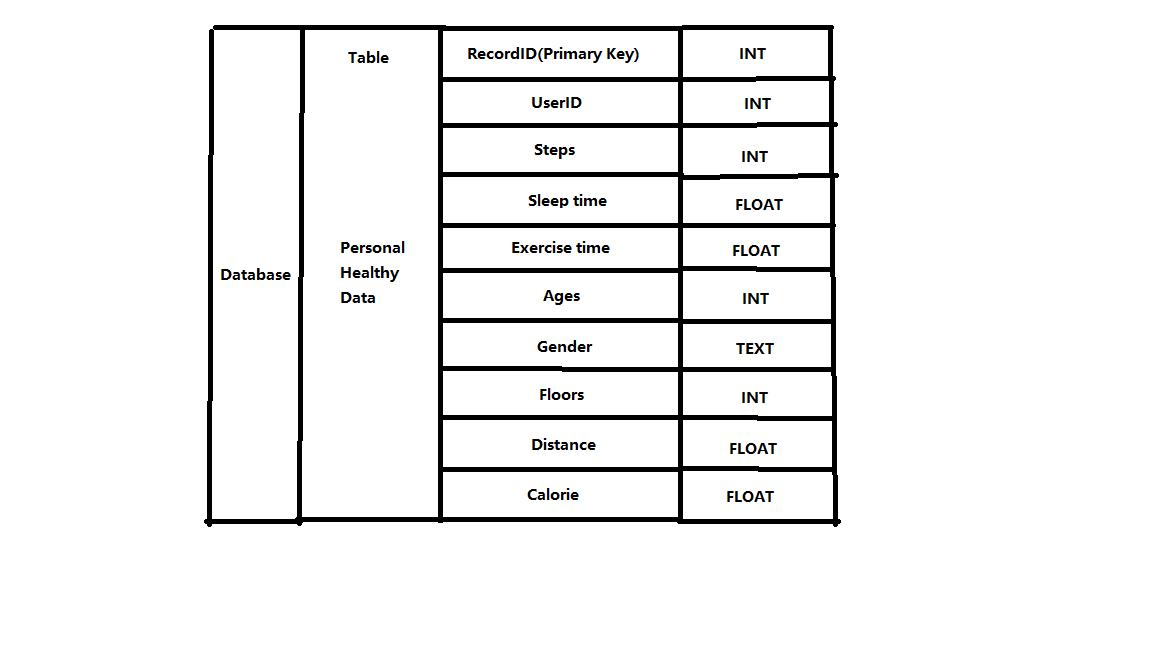


Figure 1 Data structure of Healthy Data

Design of Tests

1. **Class tests**

**Goal:** To check whether each class in the whole system can work separately, the aim is to ensure that all the basic functions can work correctly:

* The creation and accessibility of MySQL database
* The connection of MySQL database with data downloaded from the website and display the data on the web page
* Test whether the webpage can be shown correctly
* Test whether the user can register and log in
* Test the connection of social network and the return of information from it
* Test the user dashboard functions
* Test whether the analyst can view data correctly
* Test whether all graphs or Line Chart can be displayed correctly
* Test the getting advice function
* Test whether users can custom their diet plan

**The results:**

* The MySQL database is created successfully in our computer.
* The data can be stored in database and we can download or upload and modify these data from our remote devices.
* The UI display context correctly as we design it.
* Users can sign up and all the profiles registered can be saved in database.
* Connection of social network is successful and we can get the information we needed.
* The activity condition information can be displayed on the dashboard.
* Analysts can see given user’s healthy information and can give advice to the user.
* Different kinds of line charts can be shown based on user’s health condition.
* The request of getting advice can be processed correctly.
* The user custom diet plan can be loaded in webpage and can be modified by user.

1. **Functional unit tests**

**Goal:** To test the running condition of all the central parts of the application: the data analysis and interactions between different parts

1. Test ID: TC1\_Register

Unit to test: User Register

Assumption: The application is run on the user register screen, and is waiting for the user’s action.

|  |  |  |  |
| --- | --- | --- | --- |
| **Input Requirement** | **Expected Output** | **Pass/ Fail** | **Comment** |
| User’s information | Register successfully | Pass if the register is successful. | This is to make sure the user’s information is correct |
| Wrong information or information is incomplete | The error information | Pass if the system can detect the error |

1. Test ID:TC2\_Login

Unit to test: User Login.

Assumption: The application is run on the login screen, and is waiting for the user’ s action.

|  |  |  |  |
| --- | --- | --- | --- |
| **Input Requirement** | **Expected Output** | **Pass/ Fail** | **Comment** |
| Vail UserID and Password | User log in successfully | Pass when the user log in and show correct information | This is to make sure that user can access their profile successfully. |
| Incorrect userID or password | Error information | Pass when system shows the log in is rejected |

1. Test ID: TC3\_DataCollection

Unit to test: Healthy data collection unit

Assumption: The user has logged in and at the user main page, and is waiting for collecting user’s data.

|  |  |  |  |
| --- | --- | --- | --- |
| **Input Requirement** | **Expected Output** | **Pass/ Fail** | **Comment** |
| The user information | All the data related to user | Pass if we can get the data we wanted | This is to make sure we can get healthy data from Monitor device’s API |
| Invalid information | Error message | Pass if it displays the error info. |

1. Test ID: TC4\_Dataanalysis

Unit to test: Data analysis unit for analyst

Assumption: The user has requested data analysis and the analyst receives the request.

|  |  |  |  |
| --- | --- | --- | --- |
| **Input Requirement** | **Expected Output** | **Pass/ Fail** | **Comment** |
| Start Data analysis | Show the user’s healthy data to analyst | Pass if the corresponding information is shown correctly | This is to make sure the analyst part is working well. |

1. Test ID:TC5\_ConnectSocialNetwork

Unit to test: Social Network connection unit

Assumption: The user has logged in and at the main page, and sends request to connect to social network

|  |  |  |  |
| --- | --- | --- | --- |
| **Input Requirement** | **Expected Output** | **Pass/ Fail** | **Comment** |
| Start connection to social network | All the data related to user | Pass if we can get the data we wanted | This is to make sure that we can connect to social network to retrieve data and share on it. |