

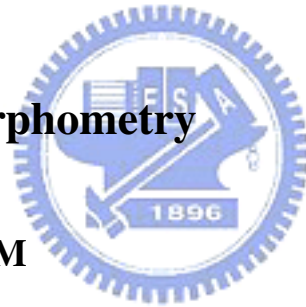
Chapter 2

Feature Selection and Extraction



This chapter is concerning how we decide features of subjects for classification. As proved correlation between diseases and the brain volume, observations of the brain volume are considered as good features to differentiate between normal subjects and abnormal ones. Presumably, it is intuitive that every unit of brain structures is available as a feature. However, the size of a medical image is $256 \times 256 \times 124$ and it costs enormous computations to analyze all images if each unit is taken as a feature. In addition, inclusions of all units may contain both of good and bad features and cause a poor classification outcome. So, it is expected that better features are extracted to obtain a better result. In this work, we use voxel-based morphometry technology to gain better features and select more useful subset of extracted features with principal component analysis.

2.1 Voxel-Based Morphometry



2.1.1 Introduction to VBM

Voxel-based morphometry (VBM) is a neuroimaging analysis technique that allows investigation of a voxel-wise comparison of the concentration or volume of brain tissues between two different groups of MR images [10]. It was first proposed by Wright *et al.* in 1995 for characterizing regional tissues' differences in structural MR images [11]. Since then, many studies of brain morphometry have been presented, for instance, schizophrenia [12], attention-deficit hyperactivity disorder(ADHD) [13], Alzheimer's disease [14] and bipolar disorder [15]. Furthermore, these studies often obtained some particular morphometric measurements from known and definite brain regions and had lots of findings related to those specific measurements.

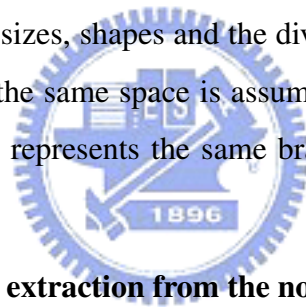
VBM has been widely used for brain structure studies because of its simplicity and feasibility. Moreover, it is a whole-brain technique that means scientists have no need to know where differences between two groups may exist and have a more clear sight of

disease pathologies. In other words, it is unbiased to any particular brain structure and gives a impartial and thoroughgoing assessment of anatomical differences everywhere in brain [10].

The basic VBM concept involves five steps described in order as below [10]. Figure 2.1 is the flowchart of the basic VBM.

1. Spatially normalization of all the images to the same stereotactic space

As shown in the figure 2.2, the spatially normalization is to register all subjects' MR images obtained by a scanner into a standard stereotactic space defined by a template image. It is impossible to compare each voxel of different MR images in native space because of different sizes, shapes and the diversity of the scanning position. So, the transformation into the same space is assumed that every voxel of different images in normalized space represents the same brain tissue and could be compared with others.



2. GM, WM and CSF extraction from the normalized images

The spatially normalized images are segmented into several different tissue classes which are grey matter (GM), white matter (WM), cerebrospinal fluid (CSF) and other nonbrain partition. A modified model cluster analysis method is often used for tissue segmentation according to voxel intensities of images.

3. Smoothing

The normalized, extracted images of different tissues are smoothed using an isotropic Gaussian kernel for the following analysis. The smoothing step enables the data to be more closely to Gaussian field model by central limit theorem and improves the validity of making inferences about statistical analysis. Also, it atones for the vague results due to the spatial normalization.

4. Voxel-based statistical analysis for localization

The normalized, segmented and smoothed images including GM, WM and CSF are

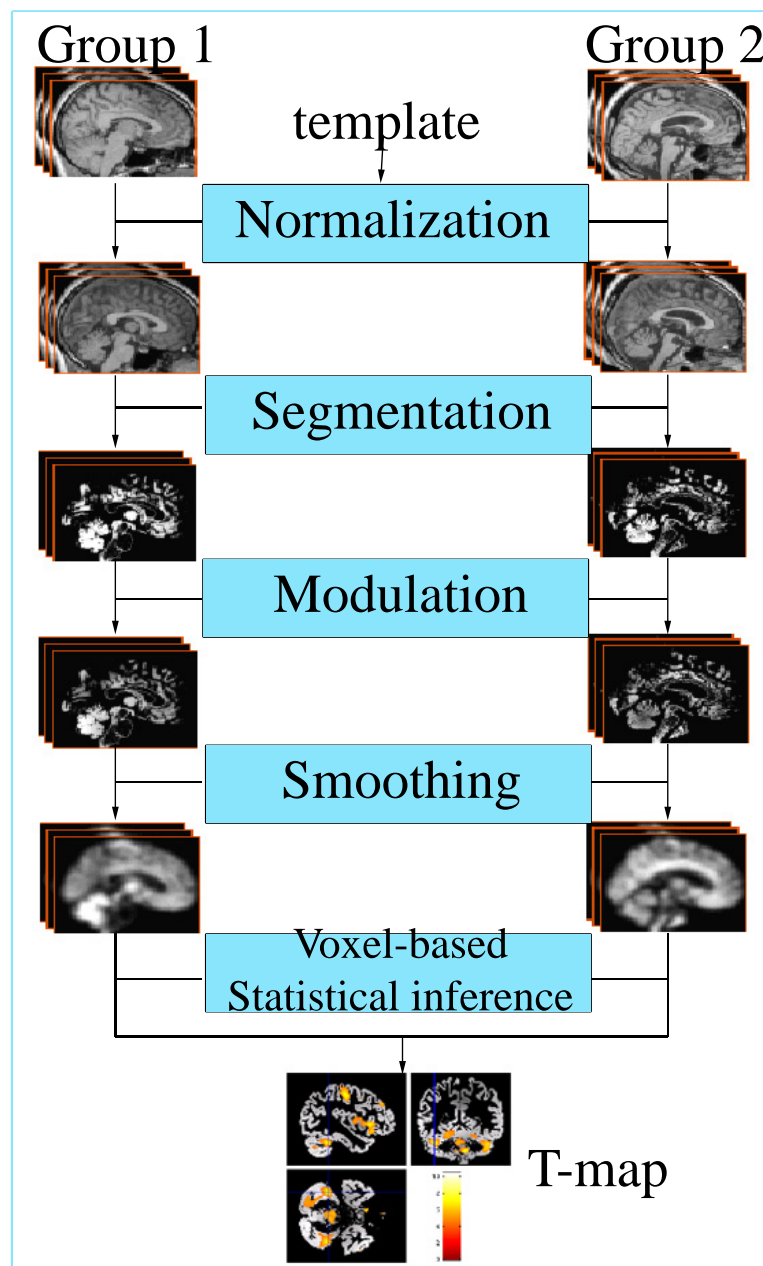


Figure 2.1: **Flowchart of basic VBM steps.** Raw images are first normalized to a standard space with a template. Secondly, GM, WM and CSF are segmented from the normalized images. Thirdly, the normalized and segmented images are modulated to correct volume changes due to previous normalization. Fourthly, the modulated images are smoothed with an isotropic Gaussian kernel to make the data close to normal distribution. Finally, a voxel-based statistical inference is applied to those normalized, segmented, modulated and smoothed data.

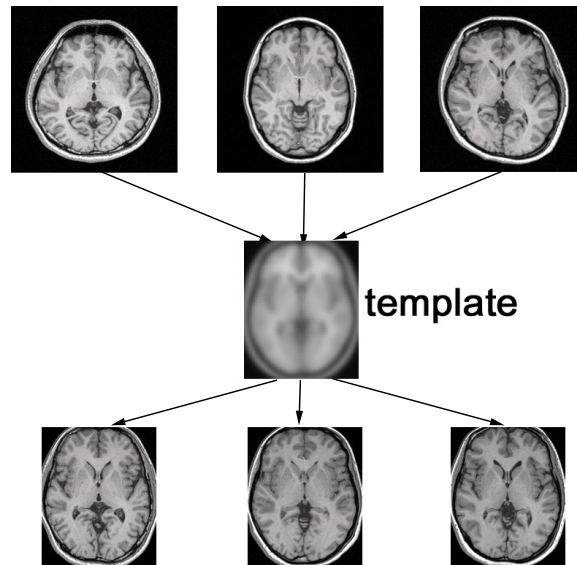


Figure 2.2: **Concepts of the spatial normalization.** Due to the different shape and size of each subject, it is impossible to compare different subjects in the native space. Thus, all images are registered into a standard space defined by a template such as ICBM152. After normalization, each subject is in the same space and can be compared.

statistically analyzed with the general linear model [16] to identify regionally discrepancy between different groups. Standard parametric statistical procedures such as t tests and F tests are often used to test the hypotheses at every voxel of brain structures to examine where are significantly related to different groups in statistic.

5. Making inferences about group differences

As a result of the fourth step, it is easy to know where different groups have diversities. In other words, a voxel is considered as a position of diversity between different groups if the voxel reaches the significant level in statistical tests. Hence, a 3D volume data with statistical parameters is obtained and we can make inferences about group differences with the data.

There are many ways to implement voxel-based morphometric analysis in terms of improvements in inferences about group differences. Some change the order of the first

three steps and some promote the normalization and segmentation techniques. In next section, one of the most popular implementations, the optimized VBM protocol [17], will be introduced.

2.1.2 Optimized VBM protocol

There are some discussions about imperfect registration influences on VBM [18]. Defective registration may lose the original image information and cause faulty segmentations. Missegmentation also leads to inappropriate comparisons between dissimilar brain structures. So, a good result of the registration will bring about a good outcome of the segmentation and lead to better interpretations of structural differences discovered by VBM.

The optimized VBM proposed by Good *et al.* [17] used a recursive method of the segmentation and the normalization to improve fine consequences of the preprocessing. The optimized VBM protocol is described as follows and its flowchart is shown in Figure 2.3 [19].

- 1. Creation of customized T1 template and a prior probability maps of GM, WM and CSF**

A customized template is created in order to reduce the deviation caused by the scanner and to be more close to the population sample to minimize distortions due to the spatial normalization. All images are normalized to ICBM 152 template (Montreal Neurological Institute), segmented into GM, WM and CSF images and then smoothed with an 8mm full-width at half-maximum (FWHM) isotropic Gaussian kernel. Each normalized, segmented and smoothed T1/GM/WM/CSF images are separately averaged to construct T1/GM/WM/CSF templates respectively.

- 2. Segmentation and extraction of affine-registered whole brain images**

The purpose of this step is automatically to remove scalp tissue, skull and non-brain

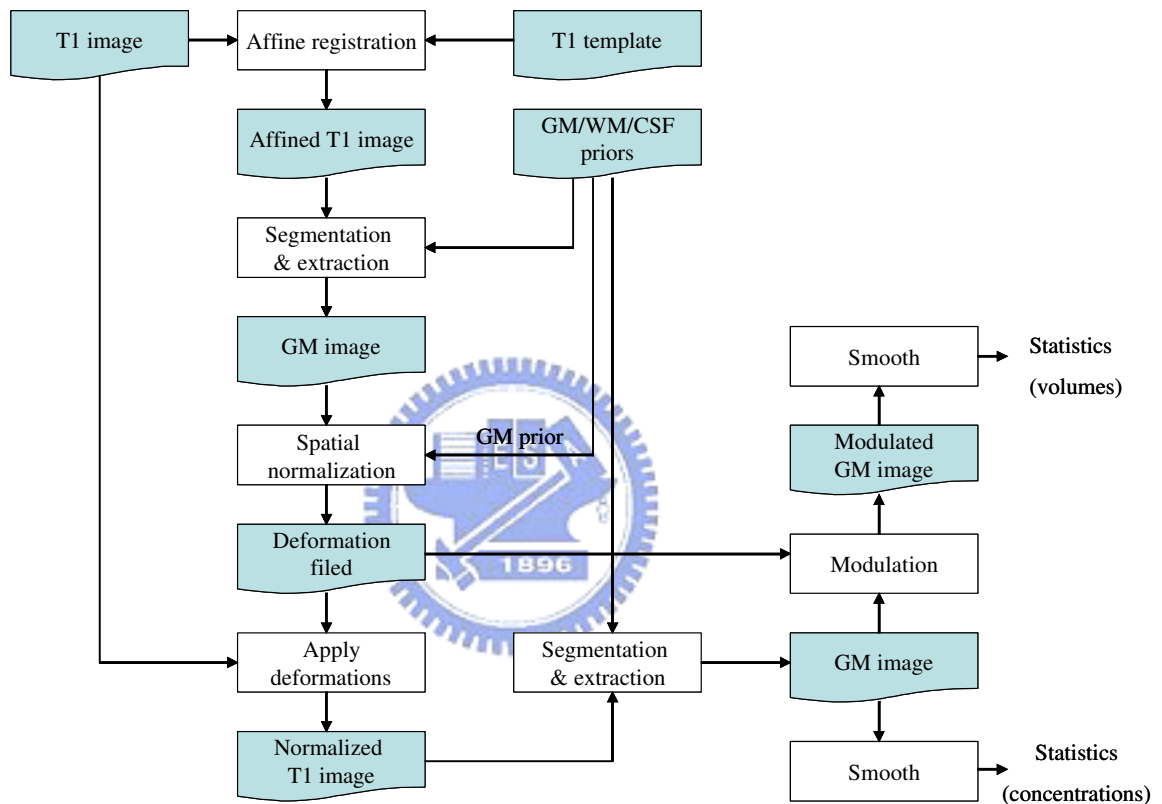


Figure 2.3: **Flowchart of optimized VBM protocol.** Optimized VBM protocol consists of the following seven steps: (1) creation of customized T1 template and a prior probability maps of GM, WM and CSF, (2) segmentation and extraction of affine-registered whole brain images, (3) obtaining optimized normalization parameters by normalizing GM/WM/CSF images into the GM/WM/CSF template, (4) normalization of whole brain T1 images with optimized normalization parameters, (5) segmentation and extraction of normalized whole brain images, (6) modulation (if need), and (7) smoothing.

tissues. Initially, GM/WM/CSF images from the original structural MR images are extracted in native space with customized GM/WM/CSF template derived from first step. Then, a series of morphological operations is applied to these segmented images to remove unconnected non-brain voxels. So, segmented GM/WM/CSF images without non-brain tissues in native space are obtained.

3. Obtaining optimized normalization parameters by normalizing GM/WM/CSF images into the GM/WM/CSF template

In this step, segmented GM/WM/CSF images without non-brain tissues acquired in second step are normalized into the customized GM/WM/CSF template individually. The process is to avoid the bias from non-brain voxels and to find the optimized normalization parameters. Then, optimal normalization parameters of GM, WM and CSF are obtained and help us measure group differences in a proper way.

4. Normalization of whole brain T1 images with optimized normalization parameters

In terms of facilitating an optimal segmentation, the optimized normalization parameters derived from the previous step are reapplied to the original whole-brain structural MR images in native space. Namely, all raw images in native space are normalized into stereotactic space. Usually, the resolution of normalized images is relatively high because of reduction of partial volume effects.

5. Segmentation and extraction of normalized whole brain images

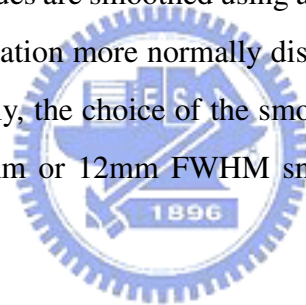
In stereotactic space, the optimally normalized whole brain structural MR images obtained from the above step are segmented into GM, WM, CSF and some other non-brain partitions. Thus, the exclusion of non-brain parts is repeated at this stage with the same methods using in second step. Finally, the optimally normalized and segmented GM/WM/CSF images are got in stereotactic space.

6. Correction for volume changes (optional)

Due to the nonlinear spatial normalization, the volumes of specific brain areas may twist and have distortions. In order to preserve the volume of a certain region within a voxel, a correction for volume changes, usually called as modulation, is incorporated. A modulated value of a voxel, represented as the volume of the voxel, is obtained by multiplying its value in the segmented images by its Jacobian determinants derived from the spatial normalization step. Besides, the unmodulated data is usually referred as the concentration of the voxel.

7. Smoothing

As the basic VBM method, the optimally normalized, segmented and modulated images of different tissues are smoothed using an isotropic Gaussian kernel. Smoothing conditions the population more normally distributed and reduces the registration error. Most importantly, the choice of the smoothing kernel is related to the expected differences. An 8mm or 12mm FWHM smoothing kernel is often used in VBM method.



8. Statistical analysis

As the basic VBM method, a voxel-wise statistical analysis is performed on those optimally normalized, segmented, modulated and smoothed images to identify regionally discrepancy between different groups. Two-sample t test is often used to evaluate hypotheses at every voxel of the brain structures. The resulting statistical parameters are formed into a 3D image, called as a t -test map. According to the t -test map, voxels with the statistical parameters reaching the significant level compose the regions representative of the detected significant group differences.

In order to have fine normalization and segmentation, the optimized VBM protocol provides an automatic brain extraction technique and construction of a separate grey and white matter templates. Moreover, it proposes a modulation step to calculate volume changes during normalization. Often, the optimized VBM protocol is implemented with SPM2

software (the Wellcome Department of Imaging Neuroscience, University College London, UK) in Matlab. In this work, we maintain concepts of the optimized VBM protocol and implement it with both of FSL (Analysis Group, FMRIB, Oxford, UK) and SPM2 softwares in Matlab 7.0 (the MathWorks, Inc. Natick, MA, USA).

2.1.3 Implementation of VBM

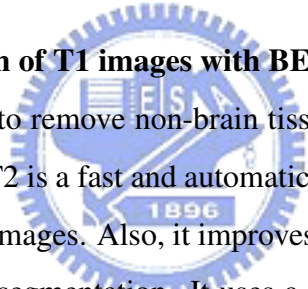
As we mentioned above, better normalization and segmentation will lead to better outcomes of voxel-based morphometric statistical analysis. Also, some studies is referred to the performance of automatic segmentation algorithms. It is shown that each kind of segmentation algorithms with optimal parameters will result in correctness of the segmentation [20]. Furthermore, there are some VBM studies using other tools to operate the segmentation instead of using SPM2 software and those researches reveal more subtle discrepancy between different groups [15]. In our thesis, hence, we tried to use both of FSL software and SPM2 software to find a better performance of the tissue segmentation.

In our experiments, we found that the efficiency of each segmentation tool depended on the quality of MR images. Our study group was collected from three different MR scanners which were set with dissimilar parameters that led to different resolution and SNR of each MR image. Surprisingly, FSL software was suitable to high quality images which have higher resolution, higher SNR and fewer artifacts, and SPM2 software was suitable to low quality, noisy images. We surmised that the phenomenon might be caused by the segmentation algorithm. In SPM2 software, images with skull and non-brain tissue were used with priors to accomplish the segmentation which was based on both of the intensity and the position. In FSL software, only images without skull and non-brain were used to perform the segmentation which was based on the image intensity. Therefore, noisy images would have a better segmentation result with SPM2 software than with FSL software. In short, we applied a suitable segmentation tool in our work according to the image quality

and then there were two corresponding implementations of VBM. One is introduced in the previous subsection and the other is as follows.

The implementation of VBM in this work is illustrated in figure 2.4. It is similar with standard VBM protocol but has different order of the normalization and the segmentation. It results from characteristics of segmentation tools. For example, the FAST segmentation could segment tissues in the native space instead of a standard space. Despite the order of the two steps, our goal is to find a good segmentation result which would solve the cross-machine problem and lead to a precise registration. Our implementation could be depicted in six steps as below.

1. Non-brain exclusion of T1 images with BET2

The logo of the FSL (FMRIB Software Library) is a circular emblem. It features a gear-like outer border. Inside the circle, there is a shield with a cross, and the text 'FSL' is prominently displayed in the center. Below the shield, the year '1896' is written. The entire logo is rendered in a blue color.

T1 images is edited to remove non-brain tissues using the Brain Extraction Tool v2 (BET2) in FSL. BET2 is a fast and automatic tool to extract the inner and outer skull and scalp from MR images. Also, it improves the contrast of the image which would achieve an accurate segmentation. It uses a surface model approach to develop the brain's surface accurately and then segments the brain images into brain and non-brain partitions [21]. There is a parameter to decide how thick the skull is in BET2. We often took the default value initially and then judged results manually for optimal parameters. All images are segmented to remove nonbrain parts in native space with BET2. Figure 2.5 illustrates the concept of non-brain exclusion.

2. Segmentation and extraction of brain T1 images with FAST

All raw T1 images without non-brain tissues derived from the previous step are segmented into GM, WM and CSF images in native space with FMRIB's Automated Segmentation Tool (FAST) in FSL. FAST also provides a function of correcting for intensity nonuniformity, also known as RF inhomogeneities. Its method is based on a hidden Markov random field model and an associated Expectation-Maximization algorithm [22, 23]. Notice that we opened the partial volume estimate option within

FAST. Thus, the value of a voxel in segmented images represents a probability of belonging to one particular tissue, ranged from 0 to 1.

3. Creation of customized GM, WM and CSF templates

All segmented GM, WM and CSF images are respectively normalized to ICBM 152 template and then averaged to construct customized GM, WM and CSF templates. These customized templates are more close to the population sample and minimize the distortion caused by normalization in later processes. The registration is operated with the normalization function in SPM2 software and the following steps are also done by SPM2 software.

4. Normalization of segmented GM, WM and CSF images

The segmented GM, WM and CSF images in native space obtained in second step are separately normalized to the corresponding GM, WM and CSF customized templates gained from the above step. So, all segmented and normalized GM/WM/CSF images of different subjects are now in the same space and could be compared. Moreover, the consequences, known as deformation field, from the normalization are stored in order to correct volume changes in next step.

5. Correction for volume changes

The purpose of this step is to restore the expansion or the shrinkage of brain volume through the normalization. We modulate these segmented and normalized images with their deformation field by SPM2 software. The value of a voxel in segmented, normalized and modulated images is considered as the brain volume in this position. In our thesis, we used these modulated images as experimental materials and probed the volume discrepancy between different groups.

6. Smoothing and Voxel-based morphometric statistical analysis

The segmented, normalized and modulated images are smoothed to be close to normal distribution. Then, a voxel-wise statistical analysis is performed on these seg-

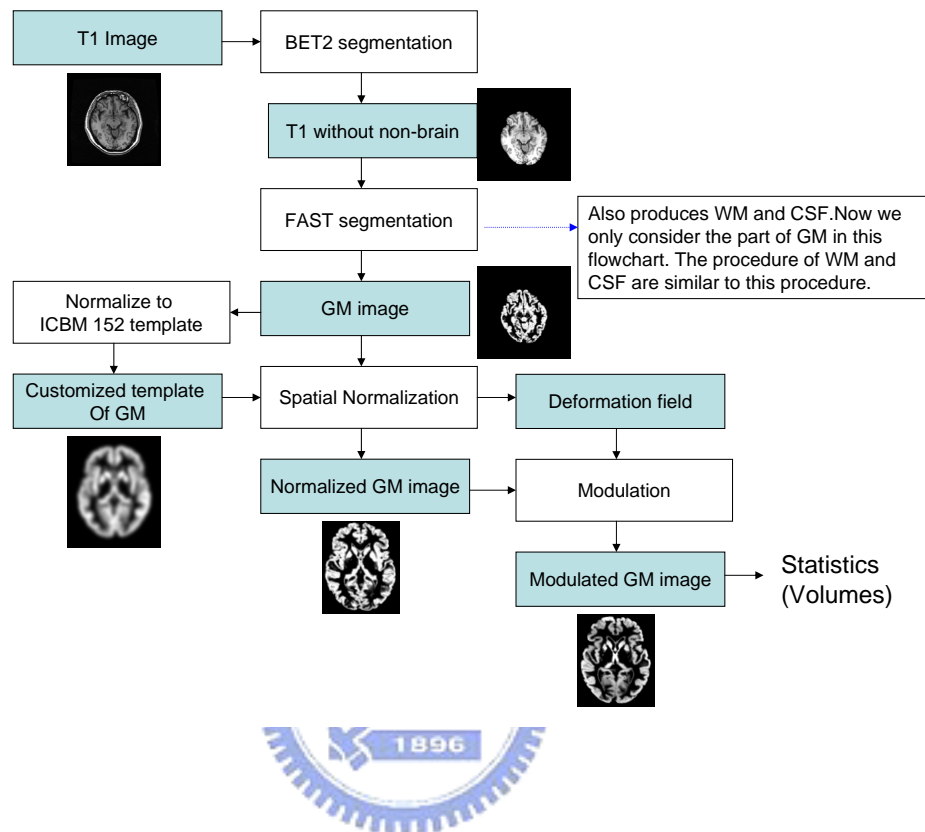


Figure 2.4: **Illustration of VBM implementation.** Our implementation of VBM uses both of FSL and SPM2 software and could be described in six steps: (1) non-brain exclusion of T1 images with BET2, (2) segmentation and extraction of brain T1 images with FAST, (3) creation of customized GM, WM and CSF templates, (4) normalization of segmented GM, WM and CSF images, (5) correction for volume changes, (6) smoothing and voxel-based morphometric statistical analysis.

mented, normalized, modulated and smoothed GM/WM/CSF images to identify regionally discrepancy between different groups. Two-sample t test was used in our work to test hypotheses at every voxel of brain structures. After the processing with SPM2, a t -test map was produced, revealed the significance of each voxel and was referred in the later processing.

VBM is used as a method of the feature selection in our system. Each voxel of the t -test map from VBM is compared with a significant level (threshold) set by human in

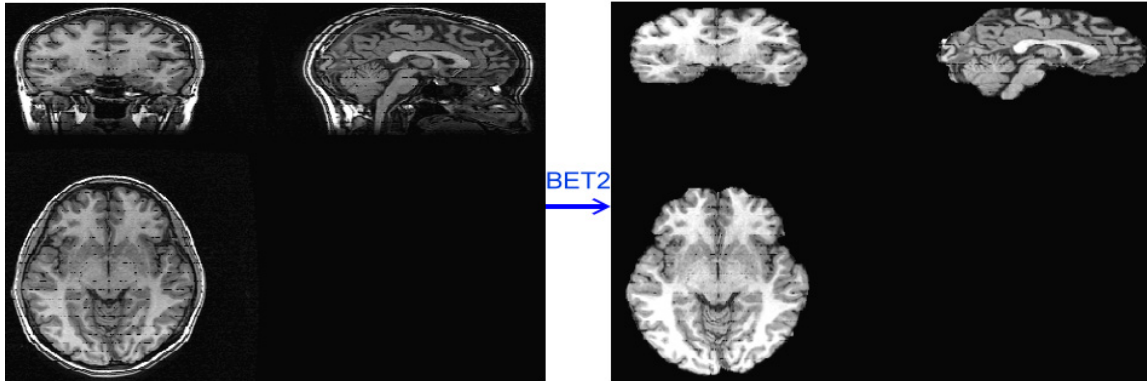


Figure 2.5: **BET2 segmentation.** This figure shows non-brain exclusion from a T1 image by using BET2. The left image is T1 images with lower contrast and the right one is an extracted brain image without non-brain materials after segmentation by BET2. The quality of the extracted image has been improved and would make a good segmentation with FAST.

order to recognize which are more important than others. If a voxel's value reaches the significant level, the voxel will be viewed as a good feature and be collected to form a new space for classification. However, those features, 146395, are large relative to the number of population sample, 91, and the dimensionality of the new space may cost a lot for computation after thresholding. In behalf of effects caused by small sample size and computation diminution, we employed a feature extraction method, principal component analysis (PCA), to reduce dimensions of the new data set and to keep the original properties of it.

2.2 Principal Component Analysis

2.2.1 Introduction to PCA

There is a well-known phenomenon in relations between the finite training data and the number of features, termed as curse of dimensionality. It was made up by Bellman in

1961 and referred to the problem that adding extra dimensions to a space would cause the exponential growth of hypervolume. In a classifier design, this problem may lead to the peaking phenomenon and affect the performance of the classifier. In practice, it is often observed that adding features may result in a poor outcome of a classifier if the number of features is large relative to that of training samples used for classification [24]. Therefore, it is important for the training data with a fixed sample size to select reliable features. In our experiments, we met the small sample size problem after thresholding and used principal component analysis to solve it.

Principal component analysis (PCA) is a linear transformation technique that simplifies a data set. It retains the characteristics of a data set in a high-dimensional space and compresses it into an improved, lower-dimensional representation for analysis. PCA, also called the Hotelling transform, was proposed by a statistician, Hotelling, in 1933 [25] and has found applications in fields such as face recognition and image compression. However, PCA is not always an optimal dimensionality-reduction procedure for classification purposes. Therefore, we will bring up a method to improve the influence on our system caused by this disadvantage in next section.

The main idea of PCA is to find most accurate data representation in a lower dimensional space. Figure 2.6 illustrates concepts of PCA. PCA determines some appropriate bases which form a subspace of dimensionality m in the original feature space of dimensionality d where m is smaller than d . All data are then projected on the subspace and represented with fewer variables. The proper subspace fits a criterion which has the minimum of projection errors. For example, the line y in figure 2.6 (b) lying in the direction of largest variance of data set produces smaller projection errors than lines in other directions and is thought as a good line to use for projection. Also, it is shown that PCA preserves largest variances in the data.

Suppose that $\{y_1, \dots, y_n\}$ represent all vectors in a set of d -dimensional samples. We want to find the most accurate representation of this set in some subspace \mathbf{W} of dimension-

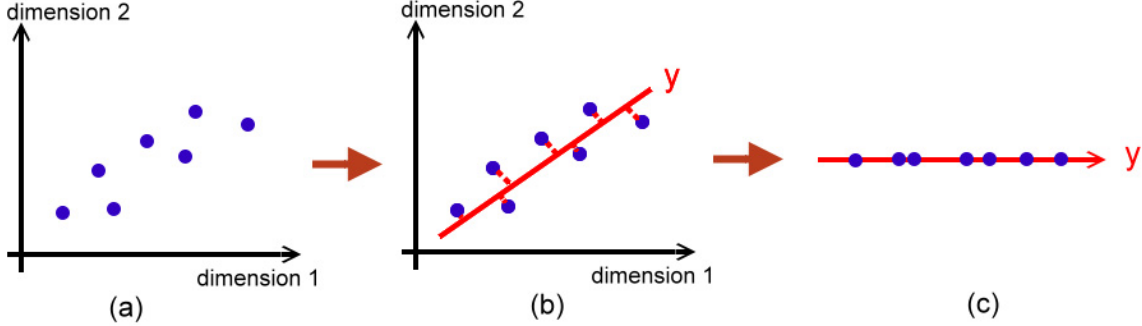


Figure 2.6: **Dimension Reduction.** Figure (a) illustrates the distribution of a 2D data set in a two dimensional space. A projection line could be found in the 2D space (figure (b)). All data points are projected to the projection line and the line represents the data set in a new and one dimensional space (figure (c)).

ality m ($m \leq d$). First, we subtract the sample mean μ from the data and the sample mean is defined as

$$\mu = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i. \quad (2.1)$$

Then, the data set is reinterpreted as $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ with zero mean. Let $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m\}$ be the orthonormal basis for \mathbf{W} . Each sample \mathbf{x}_j of the set can be written as $\sum_{i=1}^m \alpha_{ji} \mathbf{e}_i$. The total errors for representations of all the data are

$$\mathbf{J}(\mathbf{e}_1, \dots, \mathbf{e}_m, \alpha_{11}, \dots, \alpha_{nm}) = \sum_{j=1}^n \left\| \mathbf{x}_j - \sum_{i=1}^m \alpha_{ji} \mathbf{e}_i \right\|^2. \quad (2.2)$$

Thus, we can find the best base of \mathbf{W} by minimizing the total errors. After derivation of Eq. 2.2, the problem can be translated into the following equation,

$$\mathbf{J}(\mathbf{e}_1, \dots, \mathbf{e}_m) = \sum_{j=1}^n \|\mathbf{x}_j\|^2 - \sum_{i=1}^m \mathbf{e}_i^t \mathbf{S} \mathbf{e}_i, \quad (2.3)$$

where \mathbf{S} is a so-called *scatter matrix*, defined as

$$\mathbf{S} = \sum_{k=1}^n (\mathbf{x}_k - \mu)(\mathbf{x}_k - \mu)^t, \quad (2.4)$$

where μ is the sample mean [26]. The scatter matrix is just $(n-1)$ times the sample covariance matrix. Minimizing \mathbf{J} is equivalent to maximize $\sum_{i=1}^m \mathbf{e}_i^t \mathbf{S} \mathbf{e}_i$. The solution of this

problem involves the method of Lagrange multipliers and is enforced constraints $\|\mathbf{e}_i\| = 1$ for all i . Then, we find that $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m\}$ and $\{\lambda_1, \lambda_2, \dots, \lambda_m\}$ must be eigenvectors and eigenvalues of the scatter matrix, that is, $\mathbf{S}\mathbf{e} = \lambda\mathbf{e}$. In addition, eigenvalues are sorted in the descending order and eigenvectors are arranged in the order of corresponding eigenvalues. In order to minimize \mathbf{J} , the basis of \mathbf{W} must consist of the m eigenvectors of \mathbf{S} corresponding to the m largest eigenvalues. Furthermore, the larger eigenvalue of \mathbf{S} indicates the larger variance in the direction of the corresponding eigenvector.

However, it is difficult to apply PCA in practice when the sample size of data set is small relative to the space dimensionality where data set is, known as small sample size problem. Suppose that $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ represent all data in a p -dimensional samples ($p \gg n$). Directly, the scatter matrix \mathbf{S} , $\mathbf{S}_{p \times p} = \mathbf{X}\mathbf{X}^T$, is computed with the matrix size of $p \times p$. It is an inefficient way to solve eigenvalues and eigenvectors of \mathbf{S} because of exhaustive computations. Therefore, we consider the k -th eigenvectors β_k of the matrix \mathbf{T} , $\mathbf{T}_{n \times n} = \mathbf{X}^T\mathbf{X}$, with the matrix size of $n \times n$. We have

$$\begin{aligned}
 \mathbf{T}_{n \times n} \beta_k &= \lambda_k \beta_k \\
 \Rightarrow \mathbf{X}^T \mathbf{X} \beta_k &= \lambda_k \beta_k \\
 \Rightarrow \mathbf{X} \mathbf{X}^T \mathbf{X} \beta_k &= \lambda_k \mathbf{X} \beta_k, \text{ multiply } \mathbf{X} \text{ at both sides} \\
 \text{Let } \alpha_k &= \mathbf{X} \beta_k, \\
 \Rightarrow \mathbf{S}_{p \times p} \alpha_k &= \lambda_k \alpha_k. \tag{2.5}
 \end{aligned}$$

In other words, the first k -th eigenvectors of $\mathbf{S}_{p \times p}$, α_k , is a multiplier of the k -th eigenvectors of $\mathbf{T}_{n \times n}$, β_k , corresponding to the same eigenvalues. These eigenvectors are so-called the principal components (PCs). By substituting the equation, it is efficient to implement PCA in reality.

2.2.2 Feature Selection

Due to curse of dimensionality, it is often suggested that using some subsets of all features result in better outcomes for classification instead of using all features after applying PCA. The corresponding eigenvalue of each eigenvector implies the amount of variation in that direction and is usually expressed as a percentage of the total. Assume that $\{\mathbf{e}_1, \dots, \mathbf{e}_m\}$ and $\{\lambda_1, \dots, \lambda_m\}$ represent eigenvectors and eigenvalues of the reduced space after PCA. Then, the ratio of relative importance of the first k eigenvalues associated with the first k eigenvectors preserves

$$degree = \frac{\sum_{i=1}^k \lambda_i}{\sum_{j=1}^m \lambda_j} \quad (2.6)$$

of the original data information. In this work, we used two methods to select proper subsets of all features and both of them are introduced in the following.

First, principal components are selected based on variances of the data set in corresponding directions. We name this method as variance-based principal component selection. After PCA, the found eigenvalues are ranked in descending order and eigenvectors are arranged in the order of the corresponding eigenvalues. That is, the first eigenvalue is the largest among the others and preserves most of information of the data set. The smaller the eigenvalue is, the less information it preserves. Thus, it infers that the corresponding eigenvectors of larger eigenvalues should be good choices as the basis of new subspace for classification. Briefly, we select eigenvectors which have the largest corresponding eigenvalues in the remainders when the classification space is needed to expand.

Second, a method, termed as significant-based principal component selection, selects principal components for classification by their discrepancy between groups. Although the direction of the first eigenvector contains the largest variance, it does not mean that different groups of the projected data set in this direction can be differentiated from each other. Figure 2.7 illustrates this state. Therefore, we use a statistical analysis, t -test, to seek principal components where the difference between two groups reaches a significant level.

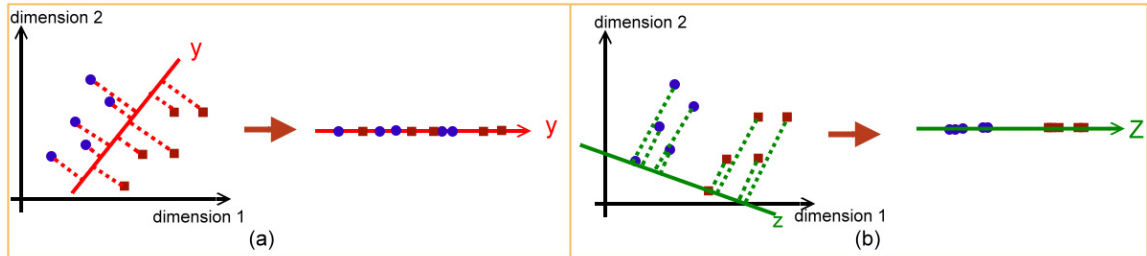


Figure 2.7: **Projection Illustration.** Figure (a) and figure (b) show two projections by using different projection lines. Although the y line in figure (a) can separate two groups, projected groups are mixed together and are bad for classification. The better choice of projection lines is like the z line in figure (b). After projection, two groups are easily differentiated from each other.

In other words, a two sample t -test would be applied to each eigenvector to compute the significance of it. Later, eigenvalues and eigenvectors are rearranged in their significant order. The more significant eigenvectors are, the more useful they are for classification. In short, we select eigenvectors which are the most significant in the remainders when the classification space is needed to expand.

In our experiments, we used both selection methods in every classification model. So, there were two classifiers, a variance-based classifier and a significant-based classifier, in each model for a specific disease. The experimental results will be shown in Chapter 4 and some comparisons of the two methods will be discussed in Chapter 5.

