Using a kernel–based approach to visualize integrated Chronic Fatigue Syndrome datasets

Ahmad Al–Oqaily **Paul J. Kennedy**

Faculty of IT,
University of Technology, Sydney,
PO Box 123, Broadway, NSW 2007,
AUSTRALIA,
Email: aaoqaily@it.uts.edu.au

Abstract

We describe the use of a kernel–based approach using the Laplacian matrix to visualize an integrated Chronic Fatigue Syndrome dataset comprising symptom and fatigue questionnaire and patient classification data, complete blood evaluation data and patient gene expression profiles. We present visualizations of the individual and integrated datasets with the linear and Gaussian kernel functions. An efficient approach inspired by computational linguistics for constructing a linear kernel matrix for the gene expression data is described. Visualizations of the questionnaire data show a cluster of non–fatigued individuals distinct from those suffering from Chronic Fatigue Syndrome that supports the fact that diagnosis is generally made using this kind of data. Clusters unrelated to patient classes were found in the gene expression data. Structure from the gene expression dataset dominated visualizations of integrated datasets that included gene expression data.
Keywords: kernel–based visualization, Laplacian matrix, data integration, biomedical datasets.

1 Introduction

Chronic Fatigue Syndrome (CFS) (Afari & Buchwald 2003) is an illness with a primary symptom of debilitating fatigue over a six month period. Currently diagnosis of CFS is generally made by clinical assessment of symptoms using a number of questionnaires or surveys measuring functional impairment, quantifiable measurements of fatigue and occurrence, duration and severity of the symptoms (Reeves et al. 2005). One goal of current research is to derive a definition of the syndrome, which goes beyond a clinical assessment of symptoms to an empirical diagnosis founded on measurements such a gene expression profiles. The motivation for this kind of research is to gain a clearer understanding of the illness and to find empirical guidelines for its diagnosis.

The question we examine in this paper is whether data visualization methods, specifically a method based on the eigenvectors of the Laplacian matrix (Shawe-Taylor & Cristianini 2004), can be used to discover patterns in biomedical datasets associated with CFS patients. Also, because there are several datasets from different sources, we are interested in creating integrated datasets and visualizing the combined data.

In the biomedical domain it is commonplace for data to be generated by high–throughput technology. One example is microarray technology (Baldi & Hatfield 2002) which generates gene expression profiles that simultaneously measure the level of expression of thousands of genes in biological samples. In general, biomedical datasets derived from high–throughput technology are described by a small number of samples (patients) and a large number of features or attributes (i.e. genes) per sample. This results in what is often referred to the 'curse of dimensionality' which makes building classifiers troublesome. For analysis of this type of data, algorithms are often applied to select features from the data to reduce its dimensionality. As a first step towards working to building classifiers for this kind of data, we initially just visualize it and look for patterns in the dataset.

In our case the data is represented by different datasets and kinds of measurements including questionnaires, complete blood evaluations and gene expression data so we also create integrated datasets by combining the individual datasets.

We apply a kernel based method to visualize the individual and combined datasets. The method (Shawe-Taylor & Cristianini 2004) we use is similar to kernel PCA (kPCA). Kernel PCA is kernel–based extension of the well known Principal Component Analysis (PCA) algorithm (Haykin 1999) and is used to reduce the dimensionality of datasets in a principled way. PCA forms a new dataset where each attribute is a linear combination of attributes from the original dataset. When the dataset is reduced to two or three dimensions it can be graphed and this allows PCA and kPCA to be used to visualize the data. Kernel PCA differs from PCA in that the data is transformed using a kernel function before the new attributes are derived. The benefit of doing this is that the attributes are not limited to being linear combinations of the original attributes and can therefore "see" nonlinear relationships in the original data. Also, using a kernel function allows visualization of non–vectorial data. We use a method based on kPCA but it differs in that whilst kPCA uses eigenvectors of the kernel matrix, the method we employ uses eigenvectors of a slightly different matrix: the Laplacian matrix. A similar approach to clustering of text with Laplacian matrices in done in (Li, Ng & Lim 2004). That approach, however, did not apply kernel functions to the data.

In section 2 we describe the datasets used for this study and the preprocessing steps applied to the data. Next, in section 3 we describe in detail the data mining and visualization approach used to identify potential patterns in the CFS datasets. In section 4 we present and results of applying the kernel based visualization method to the datasets. In section 5 we discuss these results and describe future directions for

the research. Finally, in section 6 we summarize the paper.

2 Data

In this section we describe in detail the datasets used and the preprocessing steps applied.

We used publicly available data generated as the 2006 Critical Assessment of Microarray Data Analysis (CAMDA 2006) competition datasets (CDC Chronic Fatigue Syndrome Research Group 2005). The data consists of separate datasets for patients linked by a patient identifier ("ABTID"). There were datasets with (i) survey results from fatigue and symptom questionnaires; (ii) complete blood evaluations; (iii) gene expression profiles; (iv) single nucleotide polymorphism (SNP) data; and (v) proteomics data.

The sources of data used in this study are significant because they cover the full biological spectrum from genotype through to phenotype. That is, ranging from data concerning genes through to data about their expression in the body in the form of proteins. The researchers who generated the data for the CAMDA competition hypothesize that the gene expression profile data will allow identification of "prognostic indicators" or biomarkers for diagnosis of CFS (National Center for Infectious Diseases 2005). As mentioned above, CFS is currently diagnosed using symptom questionnaires, so identification of biomarkers is potentially very significant. The analysis in this paper explores this hypothesis.

The SNP and proteomics datasets were not analyzed in this study. Analysis of the SNP and proteomics data is straightforward with our methods but will be analyzed in future studies. The SNP and proteomics data will not be mentioned further in this paper.

Data was integrated between the three datasets used in the study simply by linking of records using the patient identifier i.e. ABTID. Not all patients have data in each of these datasets. In cases where data was not available across the integrated dataset (e.g. when linking the gene expression and blood work datasets) we omitted the patients affected. This was acceptable in our situation because, as individuals in the gene expression data are a subset of patients in the clinical datasets, there was considerable overlap between the datasets and not many patients were lost.

Patients were classified into a number of categories which we grouped into three different classes: (i) those classified by physicians as suffering from Chronic Fatigue Syndrome (CFS); (ii) those classified as suffering from symptoms associated with CFS but with insufficient severity to be classified as CFS (IFS); and (iii) non fatigued individuals (NF).

In the following subsections we describe each dataset in more detail.

2.0.1 Clinical Datasets: Illness Classification and Complete Blood Work

The clinical data comprised two datasets: (i) an Illness Classification and Symptoms dataset consisting of information about patient symptoms and fatigue and (ii) evaluation of blood samples taken from patients. Each individual indicated with a patient identifier ("ABTID") has a record in both of these datasets. There were 139 CFS/ISF patients and 73 NF individuals.

The "Illness Classification SF36 MFI and Symptoms" (illness) dataset is generated based on survey results for the above mentioned patients (CDC Chronic Fatigue Syndrome Research Group 2005). The dataset includes (i) attributes that describe the general information of the patient like sex, date of birth, race, and ethnicity; (ii) the Medical Outcomes Survey Short Form–36 (SF–36) (Ware & Sherbourne 1992), as a measurement criteria for functional impairment, such as physical function, role emotional, and mental health; (iii) Multidimensional Fatigue Inventory (MFI) (Smets, Garssen, Bonke & DeHaes 1995), to obtain reproducible quantifiable measures of fatigue including "General Fatigue", "Physical Fatigue", "Active Reduction" and "Mental Fatigue"; and (iv) the CDC Symptom Inventory (Wagner, Nisenbaum, Heim, Jones, Unger & Reeves 2005) to document the occurrence, duration and severity of the symptom complex including attributes such as "Sore Throat", "Tender Nodes", "Muscle Pain, and Depression". The "Complete Blood Evaluation" dataset (blood) contained measurements of components of individual's blood as well as flags for when these measures were out of normal range.

2.0.2 Gene Expression Datasets

Microarray technology allows the high throughput analysis of global gene expression within a biological specimen. Gene expression measurements are made simultaneously for many thousands of genes. The gene expression profile of diseased cells may reflect the unique genetic alterations present and has been shown to be predictive of clinical and biological characteristics of illness for many diseases (Baldi & Hatfield 2002). A major issue in these data is the unreliable variance estimation, complicated by the intensity–dependent technology–specific variance (Weng, Dai, Zhan, He, Stepaniants & Bassett 2006). Below we describe our approach to normalizing this data. The gene expression profiles used in this study measured the level of expression of genes in blood samples from patients.

Data collected was for a subset of individuals: 118 CFS/ISF patients and 53 NF individuals. Generally there is one gene expression profile for each of these patients. A few individuals had more than one sample. The gene expression profile for a sample contains data for around ten thousand genes and data for each gene comprised around 15 attributes.

2.0.3 Preprocessing of the Clinical datasets

Most of the attributes in the questionnaire and blood evaluation datasets were used without much preprocessing.

Some attributes of the "illness" dataset, the clinical dataset containing the patient's answers to the illness questionnaires, are omitted because they are (i) skewed with almost all individuals having the same attribute value, (ii) not deemed useful for the data mining effort, or (iii) are calculated by the original researchers and would bias our efforts. The attributes concerned are "DOB", "intake classific", "cluster", "onset", "yrs ill", "race" and "ethnic". The dependent variable "Empiric" is used as the patient class and patient subtypes are combined to make three classes CFS, ISF and NF.

In the blood evaluation dataset, we add a copy of the "Empiric" attribute so that the dataset has the patient class.

All attributes of the clinical datasets apart from the patient class "Empiric" were converted to numeric values as the kernel visualisation method employed requires strictly numeric data. Binary attributes were converted to -1 and +1 for "false" and "true" respectively. Similarly, in the questionnaire dataset, categorical data values such as "mild", "moderate" and "severe" were coded to 1, 2 and 3 respectively.

Missing values were universally converted to 0. This is consistent with the coding scheme used for binary and categorical attributes. Coding of missing values to 0 is appropriate for kernel based schemes because the value 0 does not adversely effect the dot products used to build kernels. Missing values were very infrequent in the dataset and we believe that this simple approach to dealing with them is effective.

Data items in both clinical datasets were centred by subtracting the mean and attributes were normalized.

2.0.4 Preprocessing of the Gene Expression data

Data for each gene comprised a spot label (the name of the gene) and several measurements describing the level of expression of the gene as well as quality control indications of the expression measurement. We extracted the "Spot labels", "ARM Dens — Levels", "MAD — Levels" and "SD — Levels" fields. We discarded the other fields. The three statistical measures of gene expression are normalized over all arrays (samples) and patients by multiplying values with the average value of every gene over all arrays divided by the average value of every gene over the individual array.

Data for each sample was in a separate text file with filename indicative of the identifier of the patient sampled. The data for all samples was concatenated into a single gene expression file and with the patient identifier as the initial field. Additionally, the patient class "Empiric" was associated with the gene expression data through linking with the patient identifier although it was not added as an attribute to the large concatenated file.

3 Approach

The approach we have used is to visualize patients in the datasets with a kernel–based visualization method and to look for interesting features in the visualization. As shown in Fig. 1, we visualize each of the datasets (i.e. illness, blood and gene expression) in isolation, in integrated pairs (i.e. illness and blood, illness and gene, and blood and gene) and finally the integrated triplet.

As we mentioned above, the integration between datasets is done using the patient identifier. The patient class (CFS, ISF or NF) is excluded from the attributes used in the visualization because this is what we want to see in the visualization. However, each patient class is plotted with a different symbol and color in the visualization. If patients of the same class are grouped together in a visualization, this lends support to the claim that there is a relationship in the dataset. This potential relationship can be investigated in future work.

3.1 Kernel–based Visualization Approach

The approach we use is related to the kernel–based extension to Principal Component Analysis (PCA) (Haykin 1999). Principal Component Analysis is a well established method that transforms a dataset into a different coordinate system. The transformation is essentially a rotation of the dataset. The coordinates of the transformed dataset (called principal components) are orthogonal linear combinations of the original coordinates. The principal components are ordered in descending order by the amount of variance they explain in the data. Often much of the variance in the dataset can be explained by many fewer coordinates than in the original dataset (e.g. less than

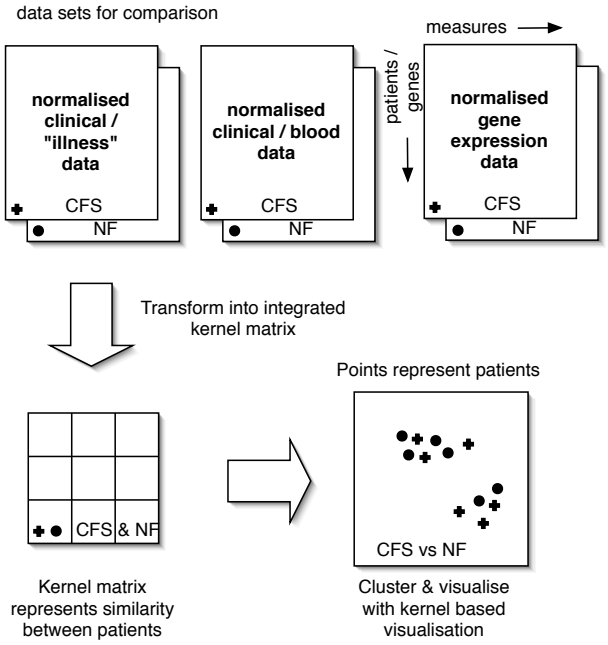


Figure 1: Approach used to visualize the individual and integrated CFS datasets.

ten). This fact means that PCA is often used for compression of data or feature selection. It also facilitates visualization of datasets by plotting the first two or three principal components of the dataset. However, as principal components are linear combinations of the original dataset, PCA has the limitation that it can only model linear relationships in the data.

There have been several approaches to extending PCA to handle nonlinear relationships. One approach is kernel PCA (kPCA) ((Müller, Mika, Rätsch, Tsuda & Schölkopf 2001), (Haykin 1999) or (Shawe-Taylor & Cristianini 2004)) which transforms the dataset $\mathbf{X}$ into a feature space using a kernel function κ before the PCA is done. Kernel PCA returns the principal components of data items in the feature space. It takes as input a Gram kernel matrix $\mathbf{K}$ which is a representation of the original dataset transformed with the kernel function. Each element $\mathbf{K}_{ij}$ of the kernel matrix is defined as

$$\mathbf{K}_{ij} = \kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \tag{1}$$

where x_i and x_j are the data items, $\phi(x_i)$ is the transformation of x_i into the "feature" space and $\langle \cdot, \cdot \rangle$ is the dot product operator. Generally it is not necessary to compute $\phi(x_i)$ explicitly. Instead, $\mathbf{K}$ is computed directly from the dataset. This is called the "kernel trick" and it means that the feature space can be very large without making generation of $\mathbf{K}$ inefficient. It also means that non–vectorial data types can be handled using special kernels such as string kernels (e.g. (Leslie, Kuang & Eskin 2004)).

Two commonly used kernel functions are the *linear kernel* and the *Gaussian kernel*. The linear kernel is defined as

$$\kappa(x_i, x_j) = \langle x_i, x_j \rangle \tag{2}$$

and is simply the dot product of the two data items. The Gaussian kernel explicitly considers the distance between data items and is defined as

$$\kappa(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{3}$$

where σ is a control parameter.

Finding the principal components in kPCA amounts to deriving the eigenvalues and eigenvectors of the kernel matrix $\mathbf{K}$. Shawe–Taylor and Cristianini describe another technique in (2004) that they say more explicitly controls the correlation between the points in the original and feature spaces. The technique is essentially the same as kPCA except that it uses (non–zero) eigenvalues and matching eigenvectors of the Laplacian matrix $\mathbf{L}(\mathbf{K})$ of the kernel matrix instead of the kernel matrix. The Laplacian matrix is defined as

$$\mathbf{L}(\mathbf{K}) = \mathbf{D} - \mathbf{K} \tag{4}$$

where $\mathbf{D}$ is the diagonal matrix with entries

$$\mathbf{D}_{ii} = \sum_{j=1}^{l} \mathbf{K}_{ij} \tag{5}$$

where l is the size of the kernel matrix.

In this study, we employ this last method using the Laplacian matrix to identify the first three principal components of datasets for visualization. We examine the data using both the linear kernel in (2) and the Gaussian kernel in (3).

3.2 Applying Kernel–based Visualization to the CFS data

Application of the kernel–based visualization scheme to our datasets was, in general, fairly straightforward. As described above, the method requires a kernel matrix representing the dataset to be visualized.

We used the linear kernel and the Gaussian kernel to make kernel matrices for the clinical datasets (illness and blood) both individually and in the integrated pair. Due to its large size, application of a kernel function to the gene expression dataset required a special approach similar to that used in computational linguistics.

Each row of the gene expression dataset represents an individual gene measurement for a particular microarray (for each patient). The straightforward approach of calculating the linear kernel matrix is to concatenate the rows of the gene expression dataset into a matrix consisting of one row for each array with a set of attribute values for each spot label ("ARM Dens - Levels", "MAD - Levels" and "SD - Levels") then to calculate the linear kernel by multiplying the matrix with its transpose.

Clearly this approach is impractical in our situation because of the large number of genes on each of the arrays. A more efficient approach, motivated by computational linguistics (see, for example, the description of generating the vector–space kernel in (Shawe-Taylor & Cristianini 2004)), for direct computation of the linear kernel matrix from the gene expression data is more appropriate.

The kernel value for two samples (i.e. microarrays) is calculated from sorted lists of genes (spot labels) associated with each array. The kernel value is calculated as the sum of the product of the attribute values for genes matching in both lists.

Computation of the linear kernel for the integrated datasets (of gene expression combined with illness and/or blood) is trivial. The linear kernel for the integrated dataset is simply the sum of the linear kernel matrices for the individual datasets.

Unfortunately this simple method of computing the linear kernel for the integrated datasets does not apply for the Gaussian kernel. In this case it is necessary to have the entire data vector. Since we never compute the data vector for the gene expression data (it is too large), we are unable to easily use the Gaussian kernel for the gene expression dataset or any integrated dataset containing the gene expression dataset.

4 Results

4.1 Visualization of individual datasets

We visualise each dataset individually. Figure 2 shows visualizations of the illness dataset. That is, the dataset containing patient's answers to the fatigue and symptom questionnaires. Figure 2a shows a visualization of the dataset with the linear kernel and Fig. 2b shows a visualization with the Gaussian kernel. The σ parameter of the Gaussian kernel was set to 100 in this case. We examined other settings for this parameter, but the value 100 gave the clearest images. As can be seen clearly in both figures, the NF individuals cluster together on the left hand side, with the ISF in the middle and the CFS patients on the right. These pleasing results are expected, as the data in this dataset is used to make the patient classifications. For the illness dataset, there is not a great deal of difference between the visualization with the linear kernel and with the Gaussian kernel.

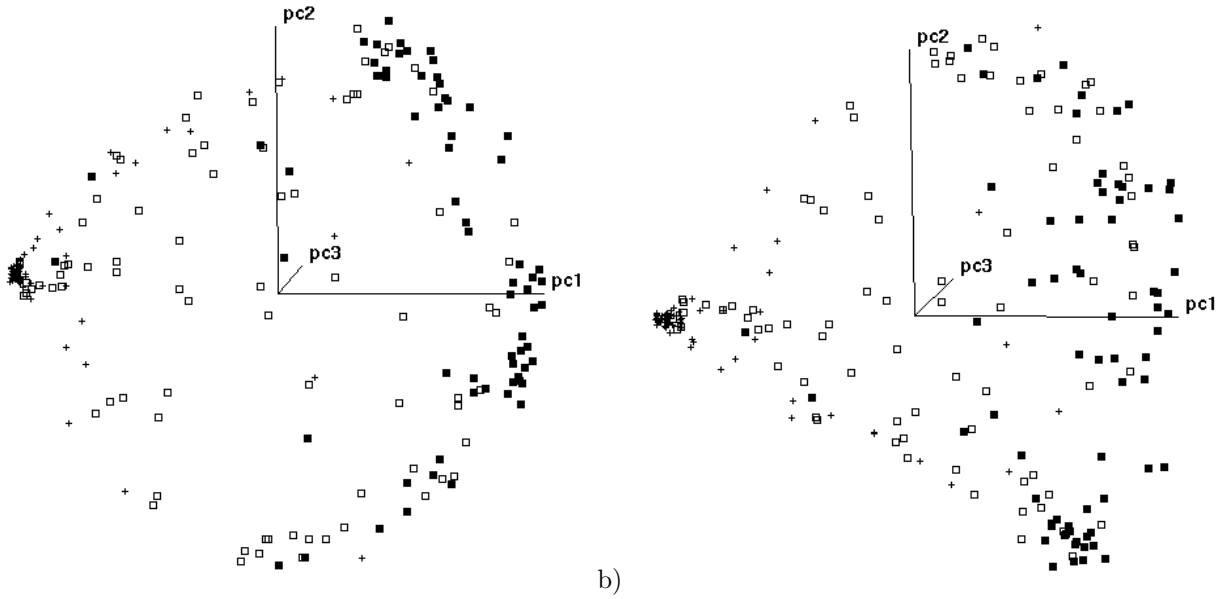
We investigated some of the patients marked as ISF that clustered with the NF individuals on the left. Patients in the dataset are actually classified with two different schemes. When we examine some of the patients that appeared to cluster incorrectly, they are classified correctly using the other scheme.

Figure 3 shows a visualization with the linear kernel of the complete blood evaluation dataset. In Fig. 4 we present visualizations of the same dataset using the Gaussian kernel. As with the illness dataset above, we tried different values for the σ parameter of the Gaussian kernel but found that the value of 100 produced the best images. Fig. 4a plots the first two principal components of the projection of the data in feature space and Fig. 4b shows the three–dimensional view. In both the linear and Gaussian visualizations of the complete blood evaluation dataset there are no clear groupings of patient classes into distinct or separable regions. This suggests to us that there are no simple "biomarkers" in the blood evaluation dataset associated with CFS. In Fig. 4a there are some small groupings of CF patients particularly five clustered in the center of the diagram that may warrant further investigation.

Finally, in Fig. 5 we present visualizations using the linear kernel for the gene expression dataset. Figure 5a shows the two–dimensional plot with the first two principal components and Fig. 5b gives the three–dimensional view. Three fairly distinct clusters of patients can be seen. However, these are not naturally associated with the classes of patient, although the top right grouping in Fig. 5a contains the least number of CF patients compared the NF individuals. The same clusters are visible in the three–dimensional plot. Although not clearly shown in the diagram (Fig. 5b), these clusters are mostly embedded in a plane with only a few data points extending further along the pc3 axis. Although the clusters do not seem to be associated with the classes of patient, it would be interesting to further investigate relationships within the clusters.

4.2 Visualization of Integrated datasets

In this section we investigate patterns in the integrated datasets. First we look at the pairs of datasets. Then we examine the combination of all three datasets. We visualise the integrated datasets



a) b)

Figure 2: Kernel visualization of illness classification (questionnaire) dataset. a) linear kernel. b) Gaussian kernel with $\sigma = 100$. Axes in both graphs are the first three principal components. + = NF individuals, □ = ISF patients and ■ = CF patients.

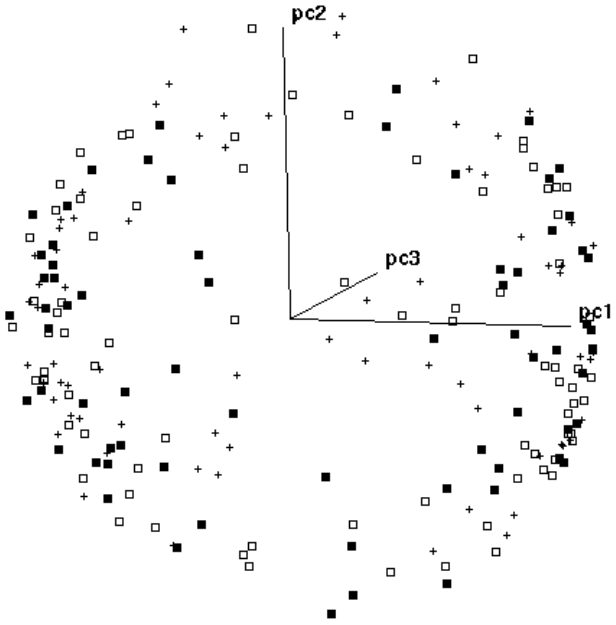


Figure 3: Kernel visualization of the complete blood evaluation dataset using the linear kernel. Axes are the first three principal components. + = NF individuals, □ = ISF patients and ■ = CF patients.

with only the linear kernel rather than both the linear and Gaussian kernel functions because (i) the use of the Gaussian kernel function did not add much to the visualizations in the previous section and (ii) because the Gaussian kernel was not applied to the gene expression dataset for the reasons described above.

Figure 6 shows visualizations of the integrated illness and blood evaluation datasets. As before, the visualization on the left (Fig. 6a) shows the two–dimensional view and on the right (Fig. 6b) the three dimensional view. It is interesting to compare these images with the visualizations for the individual datasets to see the effect of integrating the data (i.e. with Fig. 2a and Fig. 3). Recall that the NF individuals were tightly clustered in the visualization of the illness dataset but not in the blood evaluation dataset. In the integrated dataset the visualization shows the NF patients again clustered. However, instead of the compact clustering of the illness dataset, the NF individuals are now clustered in a line. In the three–dimensional visualization most of the data points are on the surface of a sphere and the NF individuals appear as line of "longitude".

Next we integrate the blood and gene expression datasets and visualise with the linear kernel function in Fig. 7. Comparing with the graphs of the individual datasets in Figures 3 and 5 it can be seen that the structure of the gene expression dataset dominates. The diagram of Fig. 7a seems to be the reflection across the horizontal of Fig. 5a. Similarly, the visualization of the integrated illness and gene expression datasets in Fig. 8 are essentially the same with the structure completely controlled by the gene expression dataset. We speculate that the reason that the gene expression dataset dominates the structure is that it contains many more attributes than either the illness or blood evaluation datasets.

Finally, in Fig. 9 we present the linear kernel visualization of the integrated triplet of datasets. Again, the situation is the same as above and the gene expression dataset dominates the visualization. There are again three fairly distinct clusters that are not naturally associated with the patient classes.

5 Discussion and Future Work

The kernel–based visualization approach allows us to integrate datasets in a straightforward way, particularly if we are content to limit ourselves to the linear kernel. This limitation, at least in the context of our study, does not seem to be onerous because the differences between visualizations of the individual blood and illness datasets with the linear and Gaussian kernel functions seemed to be relatively unimportant.

Being able to visualise integrated datasets in this way supports a "constructionist" approach to data analysis where we look for "global" patterns in the integrated dataset. The opposite method, a "reductionist" approach, looks for "local" patterns over subsets of attributes in individual datasets. An example of the latter approach in the context of this paper is

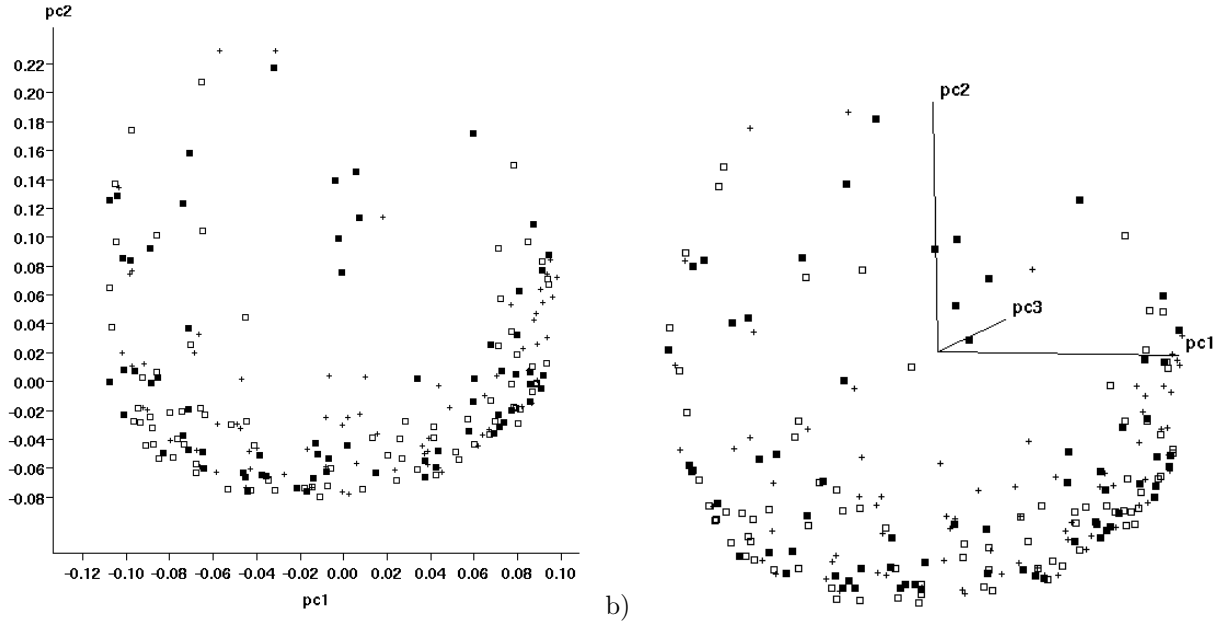


Figure 4: Kernel visualizations of complete blood evaluation dataset using the Gaussian kernel with $\sigma = 100$. a) Axes are the first two principal components. b) Three–dimensional visualization. + = NF individuals, □ = ISF patients and ■ = CF patients.

the search for small sets of genes indicative of CFS. We do not necessarily advocate a "constructionist" approach over "reductionist" ones. Rather, it is better to apply both approaches which look at different ends of the problem. A combined approach would identify global patterns, which can be used to focus the search effort for local patterns.

It was heartening that the kernel–based visualization was able to distinguish patterns in the illness data as this data was used to make the CFS diagnosis. Visualizations of the other datasets did not show such clear patterns. However, the three clusters in the datasets containing gene expression data warrant further scrutiny.

The kernel–based visualization technique is a general purpose approach and can be applied to other biomedical datasets. Indeed we intend to examine the domain of acute lymphoblastic leukaemia next. Previous work exists linking the cancer to genes so we expect to see clear clusters in this domain.

Use of specialized kernels with the technique allows visualization of non–vectorial data. We essentially used this kind of approach to efficiently build a linear kernel for the gene expression data in Section 3.2.

We are also interested in the issue of the structure from the gene expression dataset dominating over the other datasets. It is important to understand why this is the case: is it the result of the structure of the integrated dataset or is it due to the relative numbers of attributes in the individual datasets? This question must be addressed before there can be more widespread use of the technique for visualization of integrated datasets. We think that the issue here is indeed the large discrepancy between the numbers of gene expression attributes compared to the number of clinical attributes. Any method, such as the one we use, that treats attributes as equally important, will be dominated by the dataset with the larger number of attributes. One approach we plan to use to overcome this problem is to apply feature selection on the datasets *before* the data integration and visualisation. This feature selection will even up the relative numbers of attributes in the different datasets.

The linear kernel used in this study is able only to identify linear relationships in the data. Use of other kernels (such as the polynomial or Gaussian kernels) allows visualization of nonlinear relationships. In this work we were restricted to use of the linear kernel because of the size of the gene expression data. Efficient calculation of other kernels for the gene expression data is another area of future investigation.

6 Conclusion

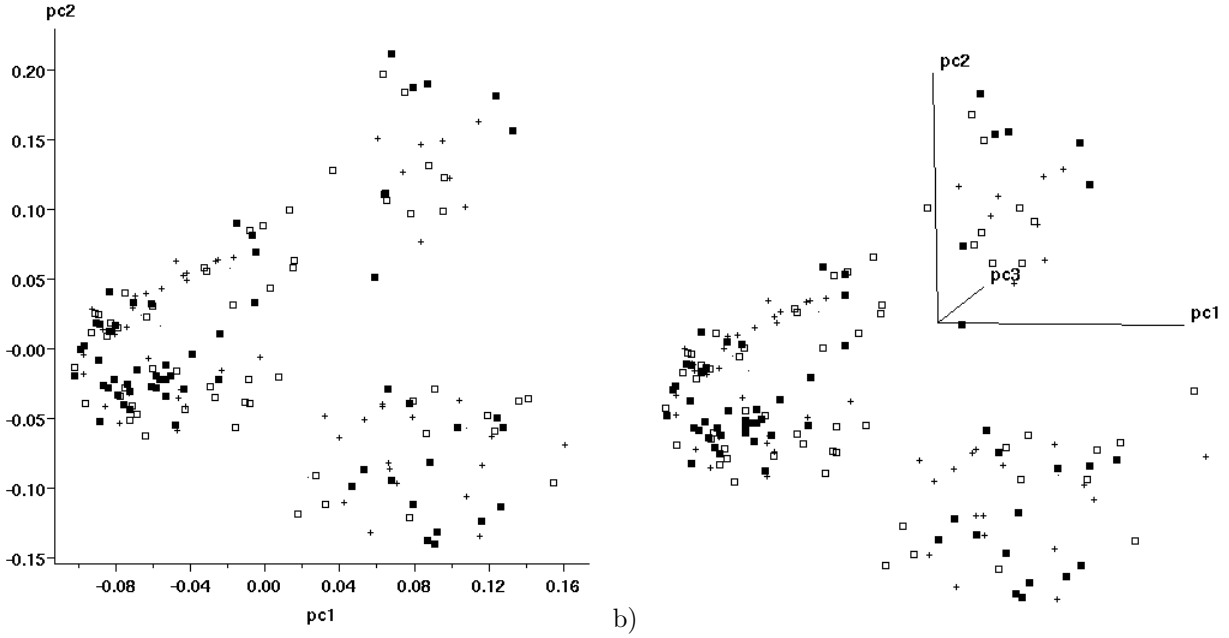
This study describes the use of a kernel–based approach using the Laplacian matrix to visualise an integrated Chronic Fatigue Syndrome dataset with symptom and fatigue questionnaire and patient classification data, complete blood evaluation data and patient gene expression profiles. Visualizations were produced for individual and integrated datasets with linear and Gaussian kernel functions. We described an efficient approach to constructing a linear kernel matrix for the gene expression data. The visualizations of the questionnaire data showed a cluster of non–fatigued individuals distinct from those suffering from Chronic Fatigue Syndrome. This observation supports the fact that diagnosis is generally made using this kind of data. The method was unable to find clusters in the other datasets that related to patient classes, although three distinct clusters were found in the gene expression data. Structure from the gene expression dataset dominated visualizations of integrated datasets that included gene expression data.

References

Afari, N. & Buchwald, D. (2003), 'Chronic fatigue syndrome: A review', *American Journal of Psychiatry* (160), 221—236.

Baldi, P. & Hatfield, G. W. (2002), *DNA Microarrays and Gene Expression: from experiments to data analysis and modeling*, Cambridge University Press.

CDC Chronic Fatigue Syndrome Research Group (2005), 'CAMDA 2006 conference contest datasets', cited; Avail-



a) b)

Figure 5: Kernel visualization of gene expression dataset using the linear kernel. a) Two dimensional visualization with axes being the first two principal components. b) Three–dimensional visualization. + = NF individuals, □ = ISF patients and ■ = CF patients.

able from: http://www.camda.duke.edu/camda06/datasets/, viewed at 10 March 2006.

Haykin, S. (1999), *Neural networks: a comprehensive foundation*, 2nd edition edn, Prentice–Hall.

Leslie, C., Kuang, R. & Eskin, E. (2004), Inexact matching string kernels for protein classification, *in* B. Schölkopf, K. Tsuda & J.-P. Vert, eds, 'Kernel methods in computational biology', MIT Press, pp. 95–112.

Li, W., Ng, W.-K. & Lim, E.-P. (2004), Spectral analysis of text collection for similarity–based clustering, *in* H. Dai, R. Srikant & C. Zhang, eds, 'Proceedings of Pacific–Asia Knowledge Discovery and Data Mining Conference (PAKDD) 2004', LNAI 3056, Springer–Verlag, Berlin Heidelberg, pp. 389–393.

Müller, K., Mika, S., Rätsch, G., Tsuda, K. & Schölkopf, B. (2001), 'An introduction to kernel–based learning algorithms', *IEEE Transactions on Neural Networks* **12**, 181–201.

National Center for Infectious Diseases (2005), 'Proposal: clinical assessment of subjects with Chronic Fatigue Syndrome and other fatiguing illnesses in Wichita', cited; Available from: ftp://ftp.camda.duke.edu/CAMDA06_DATASETS/wichita_clinical_irb_protocol.doc.

Reeves, W. C. et al. (2005), 'Chronic fatigue syndrome — a clinically empirical approach to its definition and study', *BMC Medicine* **3**(19).

Shawe-Taylor, J. & Cristianini, N. (2004), *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge.

Smets, E. M., Garssen, B. J., Bonke, B. & DeHaes, J. C. (1995), 'The multidimensional fatigue inventory (MFI) psychometric qualities of an instrument to assess fatigue', *J. Psychosom. Res.* **39**, 315–325.

Wagner, D., Nisenbaum, R., Heim, C., Jones, J. F., Unger, E. R. & Reeves, W. C. (2005), 'Psychometric properties of a symptom–based questionnaire for the assessment of chronic fatigue syndrome', *BMC Health Quality Life Outcomes* **3**(8).

Ware, J. E. & Sherbourne, C. D. (1992), 'The MOS 36–item short form health survey (sf–36): conceptual framework and item selection', *Med. Care* **30**, 473–483.

Weng, L., Dai, H., Zhan, Y., He, Y., Stepaniants, S. B. & Bassett, D. E. (2006), 'Rosetta error model for gene expression analysis', *Bioinformatics* **22**(9), 1111–1121.

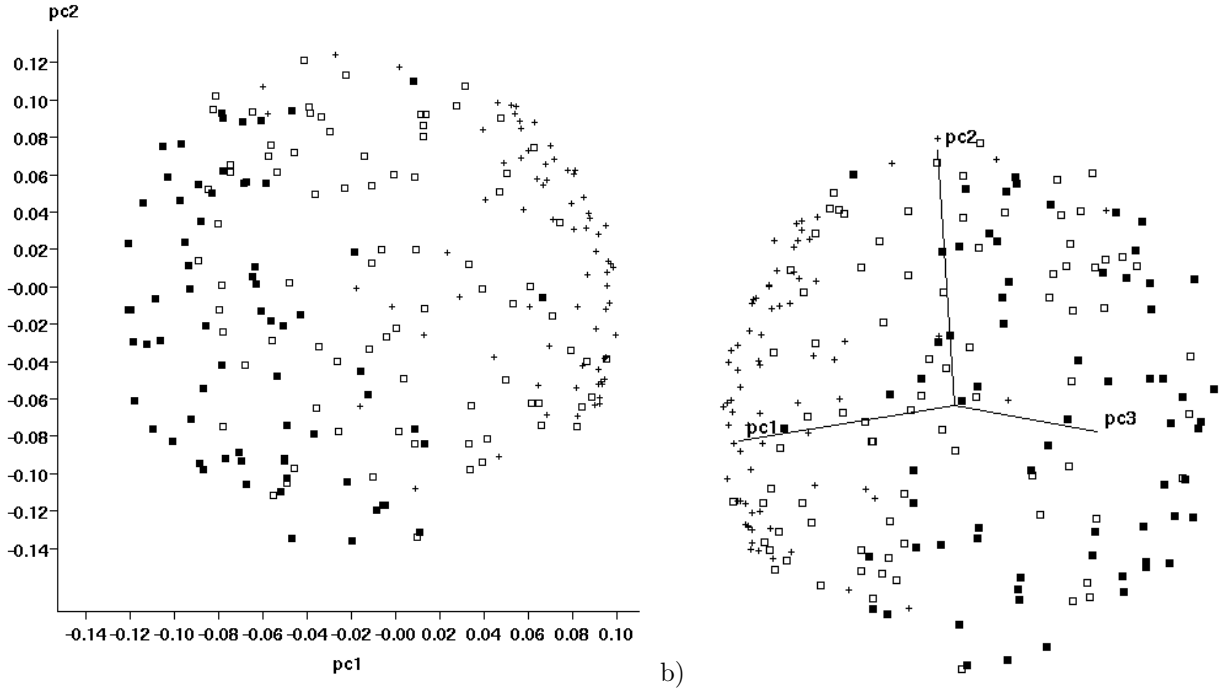


Figure 6: Kernel visualization of integrated illness and blood datasets using the linear kernel. a) Two dimensional visualization with axes being the first two principal components. b) Three–dimensional visualization. + = NF individuals, □ = ISF patients and ■ = CF patients.

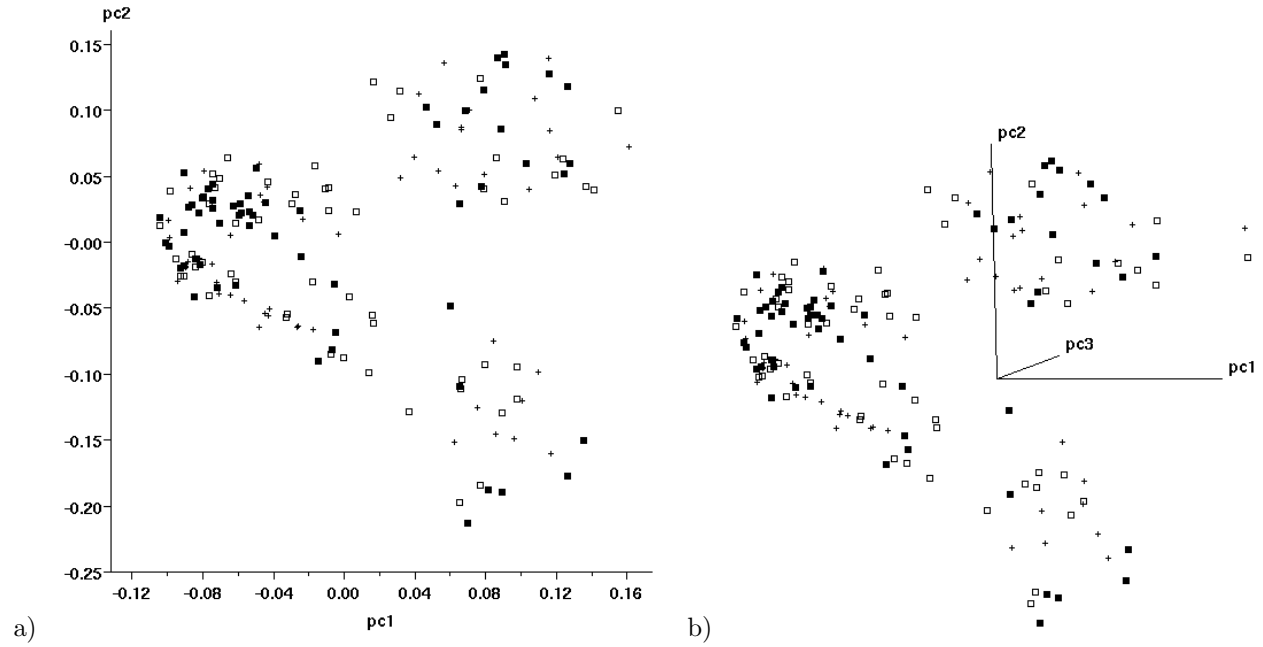
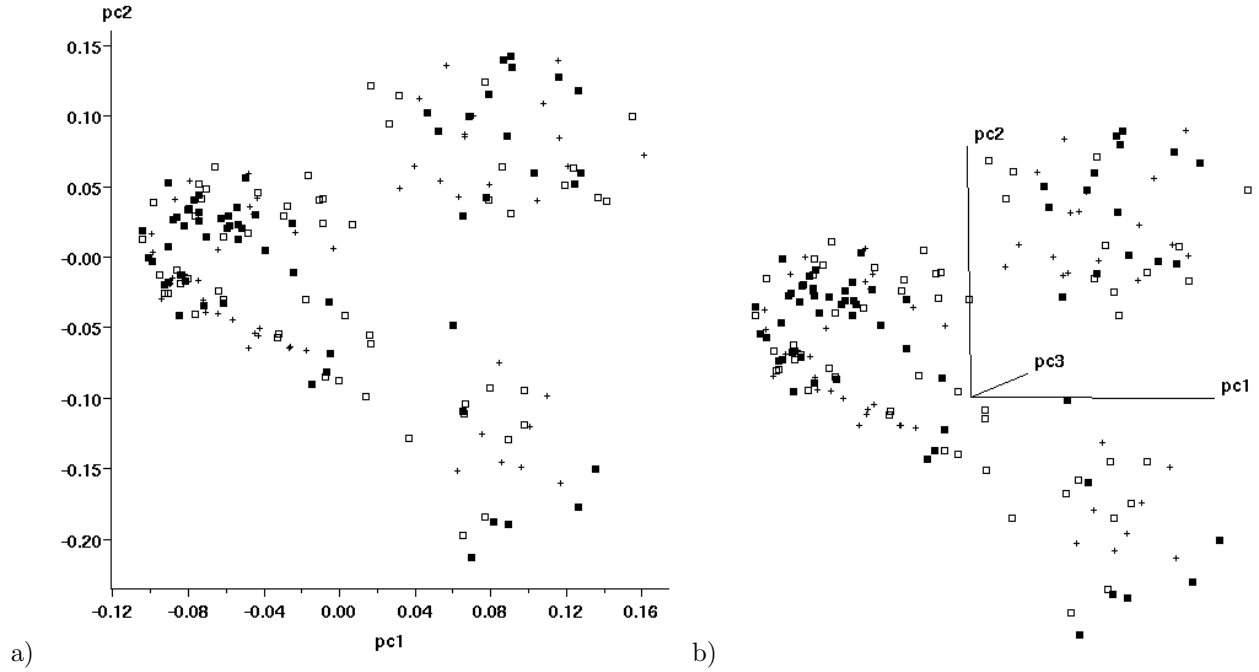
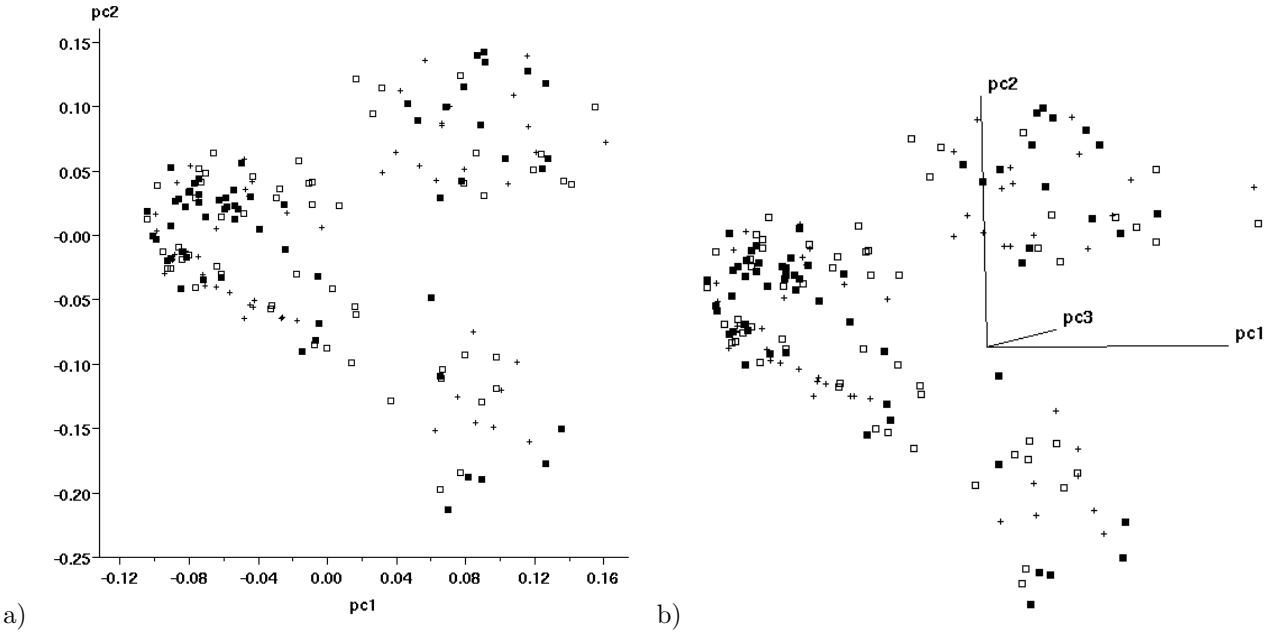


Figure 7: Kernel visualization of integrated blood and gene expression datasets using the linear kernel. a) Two dimensional visualization with axes being the first two principal components. b) Three–dimensional visualization. + = NF individuals, □ = ISF patients and ■ = CF patients.



a) b)

Figure 8: Kernel visualization of integrated illness and gene expression datasets with the linear kernel. a) Two dimensional visualization with axes being the first two principal components. b) Three–dimensional visualization. + = NF individuals, □ = ISF patients and ■ = CF patients.



a) b)

Figure 9: Kernel visualization of integrated blood, illness and gene expression datasets with the linear kernel. a) Two dimensional visualization with axes being the first two principal components. b) Three–dimensional visualization. + = NF individuals, □ = ISF patients and ■ = CF patients.