Visualisation of High-Dimensional Data for Very Large Data Sets

David Wong WONG@ROBOTS.OX.AC.UK
Institute of Biomedical Engineering, Headington, Oxford, OX3 7LF, UK

Iain Strachan IAIN.STRACHAN@OXFORD-BIOSIGNALS.COM
Oxford Biosignals Ltd., Brook House, 174 Milton Park, Abingdon, Oxfordshire, OX14 4SE UK

Lionel Tarassenko LIONEL@ROBOTS.OX.AC.UK
Institute of Biomedical Engineering, Headington, Oxford, OX3 7LF, UK

Abstract

This paper proposes a modification on the Sammon map algorithm for data visualisation. The modification, known as the Sparse Approximated Sammon Stress(SASS), allows mappings to be produced for very large data sets of the order of 10^6 points. While the technique may be useful in a variety of applications, the results presented here will demonstrate its usefulness for visualising patient deterioration in vital sign data collected from step-down unit hospital patients. A final result demonstrates an application of the SASS visualisation for drug safety analysis.

1. Background

In the field of patient monitoring in critical care, researchers are often overwhelmed with large quantities of high-dimensional data. The data typically consist of simultaneous readings of vital signs such as breathing rate, blood pressure, temperature and arterial-oxygen saturation. The new generation of automatic patient monitors and hospital IT systems enable data to be collected quickly and efficiently, so it is no longer unusual for researchers to deal with data sets containing millions of data points.

Initial exploration and analysis of such high-dimensional data is a difficult task. Any analytic tools or algorithms must deal with the data in a coherent and intuitive manner in order to provide useful insight, but must also be usable with large volumes of data.

One important aspect of high-dimensional data analysis is visualisation. This involves transforming the original data to a visualisation space with fewer dimensions. Typically, two or three dimensions are chosen so that the results can be plotted for visual inspection. The transformation is chosen in such a way as to maintain key aspects of the data distribution; for example, topology may be preserved between the dimensions.

A variety of visualisation algorithms have been proposed, including Kohonen's (1997) Self Organising Maps (SOMs) and kernel Principal Component Analysis (PCA) (Schoelkopf et al., 1997). SOMs use a neural network to map data onto a 2D grid such that similar data (i.e. data close to each other in the original high-dimensional space) are grouped together on the grid. This provides insight into the spatial relations within the data. In kernel PCA, the appropriate choice of kernel allows the data to firstly be mapped to a higher dimensional space so that a standard PCA in kernel space has the effect of producing a non-linear mapping between the original data space and visualisation space.

One popular alternative to these methods is the Sammon Map algorithm (Sammon, 1969). This produces a mapping which attempts to keep the Euclidean distances between all pairs of data points in the 2-D visualisation space as close as possible to those in the high-dimensional data space. Mathematically, this is equivalent to minimising the so-called Sammon STRESS objective function for N data samples:

$$STRESS = \frac{1}{\sum_{i=1}^{N}\sum_{j>i}^{N} d_{ij}^*} \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{(d_{ij} - d_{ij}^*)^2}{d_{ij}^*}$$

where the Euclidean distances between patterns i and j in the data space are denoted by d_{ij}^*, and the corre-

Appearing in *Proceedings of the Workshop on Machine Learning for Health Care Applications, 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

sponding distances in visualisation space are denoted by d_{ij}. The objective function is minimised by a gradient descent technique that adjusts the position of the points in visualisation space.

Unfortunately, there are two major drawbacks to the method. Firstly, the process of creating a Sammon Map is intractable for large data sets, as the STRESS calculation involves order $O(N^2)$ point comparisons. On a typical desktop PC, a few thousand data vectors is the practical limit. Secondly, the Sammon Map cannot accommodate new data, and must be retrained each time.

A number of authors have attempted to circumvent these problems. For instance, the Neuroscale algorithm developed by Lowe and Tipping (1997) uses a neural network trained on the data to derive an explicit non-linear transformation between data space and visualisation space that allows new points to be visualised using the interpolation properties of the trained neural network. However, this method also suffers from the same drawback of being unsuitable for large data sets, necessitating either a sub-sampling of the data used for training, or pre-clustering to a smaller set of exemplar vectors using a clustering algorithm such as k-means. At present, the authors are unaware of any method described in the literature that creates a true Sammon map for large ($> 10^4$ point) data sets in reasonable time.

2. Method

We propose a novel alternative to the original Sammon Map algorithm which we have named the Sparse Approximated Sammon STRESS(SASS). SASS reduces the problem to one of order $O(N)$ by sub-sampling from the complete set of inter-point distance pairs to approximate the Sammon STRESS. In practice, it has been discovered that many of the inter-point distances can be removed from the STRESS calculation, with little effect on the Sammon Map output. The method used to sub-sample is critical for obtaining an accurate mapping and is discussed further in the following section. Formally, if we define S to be a sparse subset of the index pairs (i, j) for which the Euclidean distance is calculated, then the modified STRESS objective function to minimise is:

$$SASS = \frac{1}{\sum_{i,j \in S} d^*_{ij}} \sum_{i,j \in S} \frac{(d_{ij} - d^*_{ij})^2}{d^*_{ij}}$$

For very large data sets consisting of at least $N = 10^6$ points, a sparse distance matrix with an average of 50 distance comparisons for each point has been tested and shown to work successfully. In this case, only one distance comparison is computed using SASS for every 20,000 comparisons calculated for the original STRESS. By reducing the computational complexity in this way, the initial problem of large data sets is overcome. Furthermore, data storage is reduced by using memory saving techniques for sparse matrices. Further increases in speed are made by using an efficient optimisation algorithm, scaled conjugate gradients, in preference to gradient descent.

2.1. Initialisation of d_{ij} in Visualisation Space

In the preliminary tests, points in the visualisation space, d_{ij}, were initialised with random values, following the precedent set in Sammon's original paper. During these tests, it was clear that as the size of the data set increases and the STRESS calculation increases accordingly, it becomes likely that the STRESS optimisation procedure will get stuck in a local minimum.

SASS can be initialised in a more principled manner by using a two-stage approach. Firstly, SASS is applied to a subset of the data to produce a preliminary mapping. In this pre-mapping, the points in the visualisation space are initialised randomly. The Sammon map generated by this process creates a sparse outline, or a skeleton, of the data and so the second stage of the initialisation is to approximately map the remaining points into visualisation space using the skeleton. In this case, the distance mapping technique introduced by Pekalska et. al. (1999) was used, which creates an explicit linear transformation between the data and visualisation spaces. This provides an approximation to the transformation created by the Sammon mapping, which is generally non-linear. The result of this process is that all vectors in the data set are initialised to the correct region of the visualisation space.

In preliminary tests on a data set with with 10^6 points, a 4800 point skeleton was created to initialise d_{ij}. The SASS algorithm was then run using the new initialisation values for d_{ij}. In general, it was found that the final SASS error was smaller than for random initialisation of the d_{ij} values, and that the optimisation stage converged in fewer iterations.

2.2. Initialisation of Subset S

The SASS method can fail when a subset of the data, by chance, only possesses inter-point comparisons within the subset. A pictorial representation of this problem is presented in Figure 1. It is unsurprising that such an initialisation results in an incorrect visualisation, as the algorithm will treat the subsets as

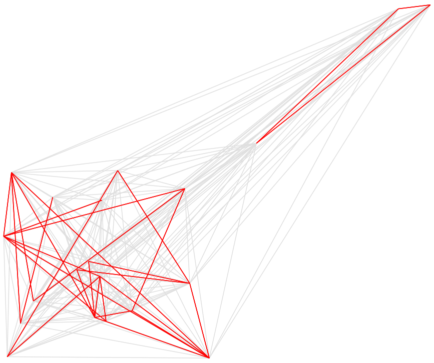

Figure 1. The graph shows an example of the connectivity between data points for the original Sammon algorithm (grey) and for the SASS algorithm (red). Each node represents a data point, and each edge represents an inter-point distance. In this example, the data has formed two unconnected subsets and SASS will fail to produce a correct mapping.

two separate data sets.

Fortunately, the probability of such an event occurring is very small. For instance, the probability of two subsets forming, where one of the subsets contains only one vector (which is equivalent to one point having no connections to any other point in the data set), is given by:

$$P(one\ point\ disconnected) = N(1 - \frac{2}{N})^{\frac{\lambda}{2}N}$$

where λ is the average number of connections per point such that $\frac{\lambda}{2}N$ is the number of elements in set S, and N is the number of data points in the whole data set, as before. For a data set with over 10^6 points and and average of 50 connections per point, the probability of one point being disconnected is of the order of 10^{-16}. To prevent this problem from occurring at all, we ensure that the connections within the data set form a minimum spanning tree. The simplest way to do this is to initially connect each data vector to its neighbours, so that the n^{th} data vector in the set of N data vectors has distance comparisons to the $n - 1^{th}$ and $n + 1^{th}$ vectors.

SASS can be further enhanced by considering the manner in which the subset S of inter-point connections is chosen. In order to test the effectiveness of alternative choices of S, a unit-cube synthetic data set was created. This consisted of a 3D unit-cube with normally distributed data at each of the corners, so that there were 20×10^4 3D vectors in total. Furthermore, the $(1, 1, 1)$ data vector was added twice to the set as two distinct data points to test whether data are mapped consistently.

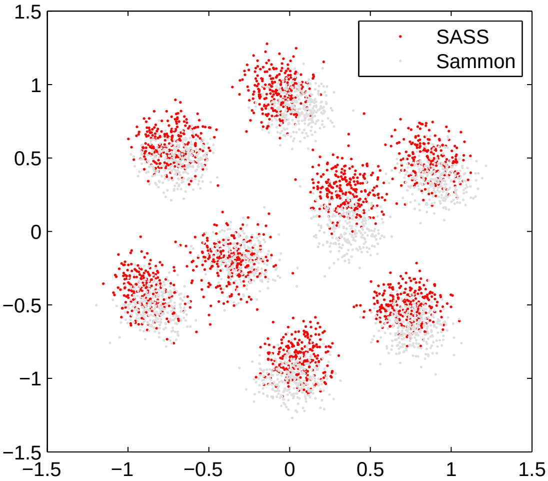

Figure 2. A Sammon Map for the unit-cube data set, containing 2×10^4 points. The SASS Sammon map is shown in red, and the output from the original method is shown in grey. In both instances, the separate data clusters are clearly visualised

In the initial tests, elements in S were chosen by selecting two data vectors at random. Figure 2 shows the result from SASS on the cube data set in red compared to results created directly from Sammon's algorithm in grey. The eight clusters corresponding to the corners of the cube are correctly mapped, and it is clear that SASS works satisfactorily. Although the results are acceptable, in order to maintain accurate local and global structure, the proportion of local and distant inter-point connections is of critical importance.

One natural way to do this is to force each data point to have an equal number of connections to both near and far points in the data set. Local and distant points can be defined for any data set as follows. Firstly, the data set is clustered using a technique such as K-means. Once the points have been grouped, half of the total inter-point connections that form set S are selected such that the two connected points are within the same cluster. These are defined as 'local' connections. The remaining inter-point connections are chosen so that any two connected points are from different clusters. Alternatively, for time series data where variation is slow compared to the data collection rate, one would expect consecutive samples to appear locally in visualisation space. Therefore, local connections can also be defined by appropriately partitioning a time series data set.

The unit-cube data set was retested using this method to define local and distant connections. Again, Figure 3 shows that the global structure was adequately captured. The duplicate points are highlighted in red, and visual inspection shows that they were mapped consistently. To quantify the accuracy of the mapping, the dataset was visualised 200 times for both a randomly initialised set S, and for the alternative method described above. In each of the 200 Sammon maps, the Euclidean distance between the mapped duplicate points was recorded, and the mean of these was calculated. For the randomly initialised set, the mean distance was 0.05, while the mean distance in the alternative method was 0.02. This indicates that it is important to ensure a sufficiently high proportion of local connections, and that selecting S at random is sub-optimal.

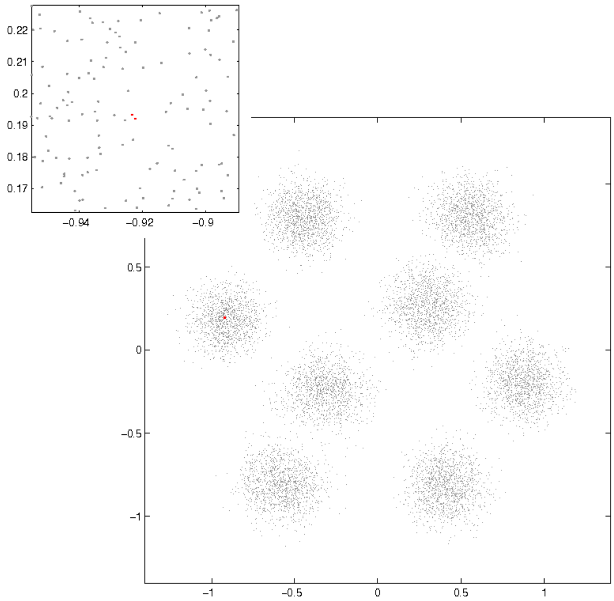

Figure 3. A Sammon Map for the unit-cube data set, containing 2×10^4 points. The SASS Sammon map is shown in grey. In the left-most cluster, the visualisation of the duplicate points at [1,1,1] are highlighted in red. The subfigure shows the left-most cluster in greater detail so that the duplicate points can be distinguished easily.

3. Results

We have used the SASS method as a tool for initially exploring extremely large data sets. Results so far have been encouraging, and have provided insight into ways of improving data fusion models for patient monitoring. The data set used to generate Figures 4 and 5 is taken from a clinical trial on a hospital step-down unit at the University of Pittsburgh Medical Centre(UPMC), and contains vital sign recordings taken over an eight week period for a total of 300 patients (Hravnak et al., 2008b). For each patient, four vital signs, the heart rate, breathing rate, arterial-oxygen saturation and blood pressure, were recorded simultaneously in a 4D data vector. In total, 961,031 vital sign vectors were recorded which corresponds to 28,782 hours of data collection.

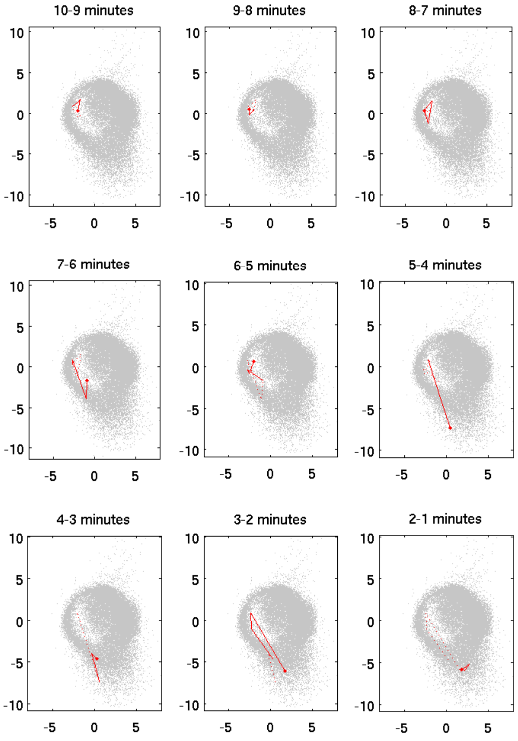

Figure 4. A time-lapse Sammon map showing the deterioration in health of patient C during the last 10 minutes of the patient's vital sign record. The points in light grey depict the vital sign distribution from the entire data set, while the points in white show the vital signs for the patient's entire stay on the ward. The lines in red mark the progression of the patient's vital signs over a one minute period.

One application of SASS allows one to see deterioration in patient health as time progresses. For the UPMC data, a time-lapse SASS map of patient C was created to depict the final ten minutes of the patient's record (Figure 4). The vital sign record for patient C is coloured in white for reference, and the entire record of vital signs recorded during the trial are plotted in light grey. Each point in the figure is a 2D representation of a 4D vital sign vector, and the vital signs recorded over one minute intervals are highlighted in red. The maps clearly show how the patient begins with relatively normal readings, which lie towards the centre-left of the population's distribution. As time progresses, the patient's vital signs become increasingly erratic as the blood-oxygen saturation readings become dangerously low. The bottom row of plots correspond to the last three minutes of the patient record where it can be seen that a number of abnormal vital signs are recorded, denoted by the points towards the edge of the grey (whole population) vital sign cluster and far away from the white (single patient) cluster, and it can be seen that there is a general trend away from normality. The fact that deterioration in patient health can be detected so clearly suggests that it is possible to use trends in time to improve patient monitoring devices.

Another SASS example is given in Figure 5. This Sammon Map depicts the vital signs for patient A and patient B from the same study in red and blue respectively. It is noticeable that the vital signs for each patient are confined to small regions of the whole distribution, indicating that there is considerable patient-to-patient variation within the bounds of vital sign normality. This is not an entirely unexpected result, as external factors such as patient age, physical fitness and reason for admission will have an effect on vital signs. However, given that in the Figure the patients' recordings do not overlap, the Sammon map provides important qualitative evidence that vital sign variation is significant enough to motivate the design of personalised data fusion models for vital sign monitoring.

A final result is presented in Figure 6, and shows the application of the SASS visualisation technique to an application in safety analysis of new drug compounds. This requires 12-lead electrocardiograms (ECGs) to be recorded from human volunteers, from which the effect of the drug on the timing of intra-beat intervals of waveform morphology are assessed. Each point on the plot represents the visualisation of the wavelet coefficients from single-beat ECG waveforms (Strachan et al., 2008; Hravnak et al., 2008a), sampled from the first eight hours of a 24 hour recording during a clinical study of the drug D-sotalol (Sarapa et al., 2004).The blue points represent the 'baseline day' where no drug was administered, and the red points represent the drug dosage day for the same subject.

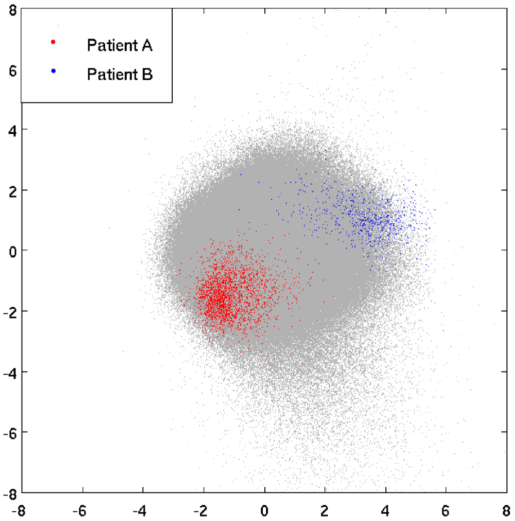

Figure 5. A Sammon Map for UPMC vital sign data. The whole data set, consisting of 961,031 4D vectors is visualised in grey. Points corresponding to vital signs for Patient A and patient B are plotted in red(left) and blue(right) respectively

The SASS visualisation was constructed from a set of 8867 beats, roughly half of which were from each day. The distance calculation for the effective Euclidean distance between two beats is more time consuming for this application because the heart rate varies, so the beats are of different lengths. Hence, before a distance calculation can be made, the beats are stretched using Dynamic Time Warping, so they lie on a common axis that minimises the Euclidean distance of the two time sequences.

The visualisation clearly shows a big effect from the drug. It is known that D-sotalol produces large changes in the morphology of the ECG wave, particular in the region of Ventricular repolarisation (T-wave). This would give rise to large differences in the Dynamic Time Warping distance measure. As can be seen, the blue (baseline) cluster is relatively compact, whereas the red (drug) points show two distinct clus-

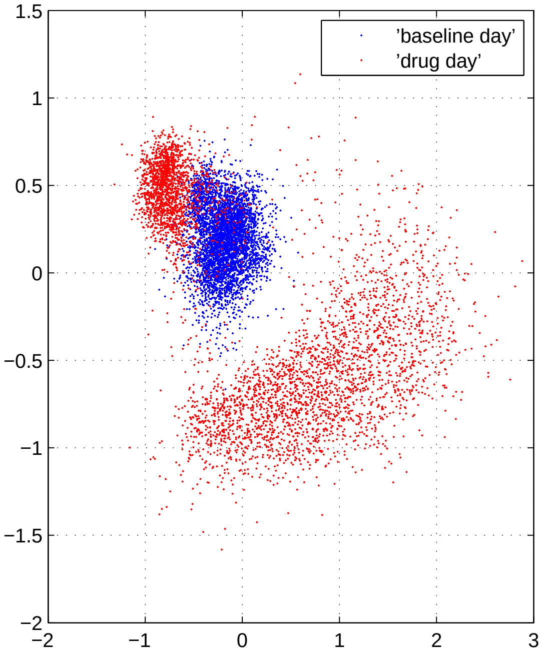

Figure 6. A Sammon map showing the effect of the drug D-sotalol on ECG waveform morphology. The 'baseline' day when no drug was taken is mapped in blue, and the red points represent readings recorded following the administration of the drug.

ters; one which is similarly compact to the baseline, and the other which is widely spread out, indicating a morphology change effect some time after administration of the dose has taken place.

The two more compact clusters are slightly displaced from each other. This is to be expected, as placements of the ECG leads can vary slightly from day to day, and this would be reflected in a small change to morphology.

4. Conclusions and Future Work

The SASS visualisation technique successfully deals with the problem of large data sets. Results in the preceding section show that sets with up to 10^6 points can be accommodated on a standard desktop PC, compared to around 10^4 points that can be mapped using standard Sammon mapping.

Visualisation of the unit-cube data set also confirms that for a medium sized data set, the SASS metric appears to be as accurate as the standard Sammon Map. The similarity between the plots in Figure 2 is encouraging, and one would expect the SASS mapping to also be accurate for larger data sets. This is not directly testable due to the Sammon algorithm limitations discussed previously.

While SASS overcomes the issue of large datasets, it continues to possess some of the other drawbacks of Sammon Maps. In particular, incorporating new data remains a problem. This is an area of current research, and we are investigating the effectiveness of methods in the literature including triangulation (Lee et al., 1977) and the distance mapping technique used previously (Pekalska et al., 1999). One promising idea is a modification to distance mapping which only assumes that local regions in data space can be accurately mapped by a linear transformation. In this way, each new data to point to be incorporated can be mapped according to its own unique, local distance map.

In the patient monitoring context, the results using the SASS technique have been especially useful for facilitating the design of 'smart' patient monitors by allowing us to compare a single patient's vital sign data to vital signs from a whole population (in a trial). Previously, such a large visualisation was computationally infeasible. Two examples have been presented. Firstly, Figure 4 showed that in some cases, deterioration of patient health through time can be clearly seen with respect to the vital sign readings of the trial population. This confirms that effective monitoring, such as the methods developed by Tarassenko et. al.(2006), can be used to provide early warning for certain adverse events and motivates the use of temporal information to improve the monitoring scheme. The second result (Figure 5), highlights the fact that, under certain circumstances, patient-specific models of vital sign data may be more appropriate than a global model of normality.

Acknowledgements

The authors gratefully acknowledge the support of the EPSRC. They are also very thankful to Oxford Biosignals Ltd. for allowing the use of the anonymised vital sign and ECG data during testing.

References

Hravnak, M., Edwards, L., Clontz, A., Valenta, C., DeVita, M. A., & Pinsky, M. R. (2008a). Defining the incidence of cardirespiratory instability in patients in step-down units using an electronic integrated monitoring system. *Arch Intern Med.*, *168(12)*, 1300–1308.

Hravnak, M., Edwards, L., Clontz, A., Valenta, C., DeVita, M. A., & Pinsky, M. R. (2008b). Impact of an electronic monitoring system upon the incidence and duration of patient instability on a step down unit. *Proc. 4th International Symposium on Rapid Response Systems and Medical Emergency Teams.*

Kohonen, T. (1997). *Self organising maps.* Springer, Berlin.

Lee, R. C. T., Slagle, J. R., & Blum, H. (1977). A triangulation method for the sequential mapping of points from n-space to two-space. *IEEE trans. on computers.*

Lowe, D., & Tipping, M. (1997). Neuroscale: Novel topographic feature extraction with radial basis function networks. *Advances in Neural Information Processing Systems 9* (pp. 543–549).

Pekalska, E., de Ridder, D., Duin, R. P. W., & Kraaijveld, M. A. (1999). A new method of generalizing sammon mapping with application to algorithm speed-up. *ASCI'99 Proc. 5th Annual Conference of the Advanced School for Computing and Image* (pp. 221–228).

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, *18*, 401.

Sarapa, N., Morganroth, J., Couderc, J., Francom, S. F., Darpo, B., Fleishaker, J., McEnroe, J. D., Chen, W. T., Zareba, W., & Moss, A. J. (2004). Electrocardiographic identification of drug-induced qt prolongation: Assessment by different recording and measurement methods. *Annals of noninvasive electrocardiology*, *9*, 48–57.

Schoelkopf, B., Smola, A. J., & Mueller, K. R. (1997). A nonlinear mapping for data structure analysis. *Lecture notes in computer science*, *1327*, 583–588.

Strachan, I. G. D., Hughes, N. P., Poonawala, M., Mason, J. W., & Tarassenko, L. (2008). Automated qt analysis that learns from cardiologist annotations. *Annals of Noninvasive Electrocardiology (to be published).*

Tarassenko, L., Hann, A., & Young, D. (2006). Integrated monitoring and analysis for early warning of patient deterioration. *Br. J. of Anaesth*, *97*, 64–68.