ElectroCardioGraphic data presented as empirical orthogonal functions: a reconnaissance

Halstead Harrison
University of Washington
Seattle, WA 98195-1640
harrison@atmos.washington.edu
Sept. 7, 2005
I. Abstract:

This reconnaissance explores the decomposition of 8-Channel ElectroCardioGraphic data into empirical orthogonal functions [EOFs], presented in a number of ways. A k-type cluster analysis separates the data files into four significant families, plus one for odds-and-sods.

More, later, I hope.

II. Data:

The data for this study were kindly made available to me by David Hadley, of Quinton Cardiology Systems, Bothell, Washington. They consist of 175 separate one-minute records, each sampled every 2 milliseconds, for 30,000 data lines of 8-columns [I, II, V1-V6] with precision of about 1%. All together, the primary data amount to about 223 MBytes.

III. Analyses:

I have condensed these data into empirical orthogonal functions [EOFs], as described in Appendix A, and grouped the files through a 'K-type' cluster analysis, as described in Appendix B. Results of this exercise are summarized in Table I, section V, and the figures of section IX.

IV. Discussion:

As an example from this effort, Figures 1a, 1c-1e, pages 9,11-13, display brief time series of the 1st four EOFs of the data file 33ic34f9, a lagged autocorrelation function, averaged waveforms, and a phase-plane display.

In 1a, the first of these curves ['EOF1', the top plot in blue] accounts for 61% of the variance, or 'information content', of all the data. It appears indistinguishable [at least to me] from a conventional analog 'QRS' strip-chart record of a normal, healthy heart [Figure 1b, page 10].
This resemblance was a surprise to me, as EOF analysis, though well understood by the 1950's, was not fashionable in the pre-digital era when electrocardiographic data processing was limited, and displays settled into the one-channel, 'PQRS' clinical mode of figure 1b. [The field is in rapid ferment today, however.] The explanation of this 'accidental' similarity, I think, runs something like this:

'Classical' ECG projects potentials measured between sets of electrodes that are intentionally placed to define left-right [x-], forward-back [y-], and up-down [z-] directions with respect to a physical torso. Subsequent operations project these potentials onto the three coordinates as \( E_z(t), E_y(t), \) and \( E_z(t) \) and sum them geometrically to their radial magnitude \( R(t) \) as

\[
E(t) = \left( E_x^2 + E_y^2 + E_z^2 \right)^{1/2}
\]

*Note that \( E(t) \) is scalar and invariant with respect to the physical orientation of the x-, y-, z- coordinates.*

'Eigenvector' ECG on the other hand, while with fancy mathematics can be expressed in terms of an 8-fold 'vector hyperspace', is more simply approached as a recipe for discovering a succession of linear combinations [Empirical Orthogonal Functions, or EOFs] of the observed potential differences [I, II, V1-V6] that are simultaneously 'orthogonal' to one another, in the sense that sums over the products of the EOFs in the form \( \sum <EOF(i,j) \cdot EOF(i,k)> \) equal zero when \( j \neq k \). This statement is equivalent to saying that EOF(j) and EOF(k) do not correlate with one another. EOF arrays have an additional and nice property that the first one, EOF1, best correlates with all of the measured potentials, EOF2 best correlates with what's left after accounting for EOF1, and so on. Thus EOF1 is 'the best' proxy for all the data, viewed at once, and a good emulator of \( E(t) \), and a classical 'PQRS' electrocardiogram.

Jumping back for the moment into vector jargon, EOF1, may be interpreted as a vector with rescaled magnitude, \( E(t) \), in a 'natural' coordinate system p-, q-, r- that may, or may not, resemble the physical coordinates x-, y-, z-.

Again, note that the magnitude of the EOF1 vector is independent of the orientation of the coordinates p-, q-, r-, but, owing to the placement of electrodes that are intentionally oriented to be sensitive to left-right,
front-back, and up-down it is a good guess that they are close to, but not identical with x-, y-, z- of a physical torso.

In figure 1a the second EOF [middle plot, green] accounts for an additional 36% of the variance, or information content of all the data. This array, EOF2, may similarly be interpreted as a vector that is perpendicular to EOF1, much in the sense that y- is perpendicular to x- in Cartesian coordinates. Because EOF2 carries information that is not correlated with EOF1, and because the latter so closely resembles 'classical' ECG displays,

there is a reasonable hope that EOF2 may add significantly to diagnostic skill, beyond that contributed by PQRS single-channel displays.

It should be emphasized, however, that EOF decompositions do NOT add information beyond that already contained in [I, II, V1-V6] data arrays.

The third and fourth EOFs [purple and rust] account for only 1.4% and 1.2% of all the variance, in this example, with only 0.26% left unaccounted by the remaining four EOFs. These may similarly be interpreted as vectors that are sequentially perpendicular to all of the preceding vectors. After EOF3, however, the geometric analogy gets harder to visualize.

Figure 1c, page 11, shows a lagged autocorrelation function of the sums of squares of the raw data [I, II, V1-V6] of the same ECG record, 331C34F9, from which may be extracted a mean heart rate and its standard deviation.

Figure 1d, page 12, shows the averaged EOFs of the same heart, with many heart beats superimposed over one apparent cycle, keyed the maxima of EOF1.

Figure 1e, page 13, shows 'phase-plane' plots of EOF1 vs EOF2 for the same heart, with color coding for the times spent in each area-increment of the plots [left-two plots] and for the signs and magnitudes of EOF3 and EOF4 [right-two plots].

---

1 The cosine transform of the lagged autocorrelation function gives the Fourier power spectrum.
Note that eigenvector analysis renormalizes the amplitudes of the EOFs to unit variance: thus the relative magnitudes of the EOFs in figures 1a are not meaningful, but the shapes of each EOF and phase relations among them retain useful information. By conventions the largest excursions of EOF1 and EOF3 are positive, and negative for EOF2 and EOF4.

It remains to be seen, however, whether such displays usefully assist diagnostic skill.

**V. Heart Rates, Irregularities, and Superpositions:**

Repeating for this discussion, figure 1d, page 12, shows strip-chart curves of EOF1 and EOF2 obtained through superposition keyed on the maximum of EOF1, as averages over all the cycles of this very regular heart [331c34f9]. When attempting similar plots with irregular ECG data, however, I ran into a family of unexpected problems. It is not at all easy with ECG data, only, to determine an average heart rate and its standard deviation for all cases, including very irregular hearts where, not rarely, the T-wave exceeds the R- [figure 1b, page 10]. The common arterial pressure pulse we feel at our wrists is greatly smoothed by the elasticity of our arteries and veins. Complex, irregular, high-frequency, and 'spikey' ECG events, however, do not map uniquely into physical heart beats. I have not yet succeeded in writing an algorithm that reliably distinguishes the latter and counts one and only one beat each for every ECG cycle, in all 175 cases.

Working around this problem, I have explored lagged autocorrelation spectra as a way to get 'most probable' heart-beat rates, with rather strenuous smoothing, but valuable phase information is lost [figure 1b]. More effort is warranted to this task, but I am aware that it may overbalance higher priorities.
VI. 'Cluster' Analyses:

I have subjected the 175 data 'objects' [individual ECG records], plus 10 'faux objects' composed of normally distributed Gaussian-random numbers, to a K-type 'Cluster' analysis, as described in Appendix B and Table I, below.

Table I
Cluster Analysis of the 175 objects of this study, together with 10 Gaussian-normal RANDOM objects.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Objects</th>
<th>Separation Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19</td>
<td>0.5765</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33FB0000</td>
<td>36A5E2A2</td>
<td>3505D73B</td>
</tr>
<tr>
<td>33F8882B</td>
<td>3506BD7C</td>
<td>3B83F5AE</td>
</tr>
<tr>
<td>35586131</td>
<td>3549E2B8</td>
<td>369E100F</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>46</td>
<td>0.6810</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>375E5C07</td>
<td>3B37BO06</td>
<td>3B8D4B61</td>
</tr>
<tr>
<td>3C4FOFO6</td>
<td>3E566F0A</td>
<td>392AA88C</td>
</tr>
<tr>
<td>387C3176</td>
<td>3B5FOCA6</td>
<td>3416DE52</td>
</tr>
<tr>
<td>3426A4C1</td>
<td>377B84C0</td>
<td>365C236A</td>
</tr>
<tr>
<td>38A1A5FC</td>
<td>3742C49E</td>
<td>3406283</td>
</tr>
<tr>
<td>3F745233</td>
<td>397DB309</td>
<td>33AECDADA</td>
</tr>
<tr>
<td>385F99CF</td>
<td>3B1D2A19</td>
<td>38EA1CBD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>44</td>
<td>0.7919</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3CCED568</td>
<td>35473BB1</td>
<td>3336DE1C</td>
</tr>
<tr>
<td>33F1C318</td>
<td>37985C0D</td>
<td>351BC93C</td>
</tr>
<tr>
<td>3AEFOF69</td>
<td>3611609</td>
<td>36793471</td>
</tr>
<tr>
<td>36556FD4</td>
<td>3B619B5D</td>
<td>356E6F60</td>
</tr>
<tr>
<td>3613B16A</td>
<td>36A8AD34</td>
<td>38FF5E48</td>
</tr>
<tr>
<td>3B9FA310</td>
<td>337E71488</td>
<td>36BA0208</td>
</tr>
<tr>
<td>36A78B18</td>
<td>38721301</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>29</td>
<td>0.9589</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3356271E</td>
<td>37F3A79D</td>
<td>34DB38EA</td>
</tr>
<tr>
<td>33CF6B2C</td>
<td>3360FD7B</td>
<td>377BAFA6</td>
</tr>
<tr>
<td>3433AECC</td>
<td>34509308</td>
<td>35F520C6</td>
</tr>
<tr>
<td>35EB651</td>
<td>39103880</td>
<td>33957C0B</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>47</td>
<td>1.1683</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[38B84037E]</td>
<td>34195739</td>
<td>3DC19040</td>
</tr>
<tr>
<td>38105307</td>
<td>33BA5197</td>
<td>344E098D</td>
</tr>
<tr>
<td>370B65EE</td>
<td>338C44FA</td>
<td>3430F8DC</td>
</tr>
<tr>
<td>33EF5B11</td>
<td>369E33EB</td>
<td>[345025F]</td>
</tr>
<tr>
<td>3357BAEE</td>
<td>36AF3E11</td>
<td>3565D0E9</td>
</tr>
<tr>
<td>RANDOM</td>
<td>RANDOM</td>
<td>33EB73ED</td>
</tr>
<tr>
<td>RANDOM</td>
<td>RANDOM</td>
<td>[344667C4]</td>
</tr>
<tr>
<td>RANDOM</td>
<td>RANDOM</td>
<td></td>
</tr>
<tr>
<td>RANDOM</td>
<td>RANDOM</td>
<td></td>
</tr>
</tbody>
</table>
In Table I, four significant clusters [with separation indexes [SI] < 1] were identified, and a 5th cluster [SI = 1.168] collected the remaining odds and sods, including the 10 random clusters. Note in the 1st cluster, which contained 19 objects with a highly significant SI of 0.576, that two objects [highlighted in blue] had been earlier identified as very closely related [Appendix B].

Similarly the 2nd cluster contains 46 objects with SI = 0.681, also quite significant. The two green objects were also earlier identified as closely related [Appendix B]. The 3rd cluster [44 objects, SI = 0.792] contains the very regular heart-beat file 331C34F9, discussed above and shown in figures 1a, 1c-e.

So far, the cluster analysis has separated 62% of the data records into significant, local clusters. The 4th cluster, however, [29 objects, SI = 0.959] is only marginally significant, but if included it brings the sorted fraction up to 79%.

The 5th cluster [47 objects, SI = 1.168] is not significantly separated. It should be commented, however, that in each of the clusters the objects closest to the local centroid are listed first [left-to-right, then top-to-bottom, as in ordinary printed text] and those farthest are listed last. Thus the 10 RANDOM objects in the 5th cluster are properly located near the last in that grouping. Their nonsense predictors, however, falsely handicap the SI index of that cluster. If they were omitted, the SI(5) index would be close to unity. In any event the earlier entries in the 5th cluster are likely also significantly grouped with one another.

VII. More Discussion:

a. Pacemakers

The data set of this study consists primarily of impaired hearts, with an apparent lack of 'normal' hearts for comparison. The record 331C34F9, however, is atypically regular, and a question rises as to whether it may have been synchronized by a pacemaker.
In a conversation with Dr. David Hadley, of Quinton Cardiology, it was suggested that a number of other ECG data records might be artificially paced. These have been identified in Table II by enclosing vertical bars, ||. Examining these files' EOF displays reveals obvious timing spikes, as in figure 2, page 14.

Note that the majority of these paced objects appear in the 5th, least-regular and least-significant cluster.

b. Limitations

The hope of this effort was to improve diagnostic power from ECG data, starting from what was perhaps an initially too naive perception that one-channel "PQRS" electrocardiographic displays were underutilizing the full information content of digitized 8-channel ["12-lead"] data. This may still be so, but with subtleties.

Continuing this effort beyond this relatively trivial reconnaissance, however, has been inhibited by lack of access to further information about the tested subjects, their diagnoses, and clinical outcomes. Several of the data records, as for example the highlighted objects in clusters 1 and 2, may be of the same patient, for example, taken at different times.

The predictors chosen for the cluster analyses were selected with arbitrary guesses, guided by numbers easy to extract, not by regression or other methods from after-the-fact information of patient outcomes. As discussed in Appendix B, their information content is equivalent to only 5 independent EOF vectors; the remaining 7 predictor combinations add only to noise. No predictors were selected from the phase information contained in the figures, which are potentially of some significance, and no correlations could be attempted between the clustering and clinical evaluations.

c. More figures.

Recognizing these limitations, it does not seem useful at this time to print figures for all 175 data files. To abstract these I have in figures 3-7, pages 15-20, plotted the 1st [and closest to the local centroid] object files of the 5 clusters. Figures 8 and 9 [pages 21-22] conclude with interesting variants.
VIII. Summary:

This reconnaissance explores the use of EOF decomposition, phase-plane displays, and cluster analyses for the presentation of ECG data. Any follow-on to this study would be greatly assisted by complementary data on diagnoses and outcomes.

IX. Figures:

"Strip-charts" of the 1st four EOFs of 331c34f9
Figure 1b:

'Classical' PQRS one-channel ECG display
Figure 1c

Lagged autocorrelation function of an apparently healthy heart, 331c34f9. The mean heart rate is 75.9 beats / min, with a standard deviation [proportional to one half the distance between the two blue dots] of 3.8 beats / min.
Figure 1d 331FC34F9

Superpositionally averaged EOFs from an apparently healthy heart.
Figure 1e  331C34F9

Phase-plane plots of the EOFs. Each heart-beat cycle circulates clockwise in these plots.

The upper-left plot shows EOF1 vs EOF2, color coded by the times spent in each area of the figure. ['hotter' colors indicate more time, cooler less]

The lower-left plot is the same as the upper-right, to larger scale, resolving details of the resting period.

The upper-right plot shows EOF1 vs EOF2, this time color coded by the magnitudes of EOF3 [hotter colors > 0 and cooler < 0].

The lower-right plot again shows EOF1 vs EOF2, this time color coded by the magnitudes of EOF4.
Figure 2a 3B84087E

Obvious brief spikes in the 1st 3 EOFs identify this case as pacemaker modulated
Figure 2b  3B8407E

Same ECG plotted more compactly. Note the nicely regular autocorrelation function but broadened superposition plot and irregular, 'smooshed', phase-plane plots. This file is leads off the 5th cluster [see Table I].
Figure 3 33FB0000

This file leads off the 1st cluster.
Figure 4  375E5C07

Lead-off ECG file of the 2\textsuperscript{nd} cluster.
Figure 5 3CCED568

Lead-off ECG file of the 3rd cluster. Note the similarity of this heart with that of 331C34F9, which is also contained in this cluster.
Figure 6  3356271E

Lead-off ECG file of the 4th cluster.
Figure 7. 3B84037E

Lead-off ECG file of the 5th cluster
Figure 8  34195739

Second object in the 5th cluster
Figure 9  33E73E5D

Towards the bottom of the 5th cluster
Appendix A: EOF analyses

Here are the crossed correlation and eigenvector matrices for the 'normal' heart data file 331c34f9:

Averages
0.325 1.174 -1.495 -1.019 -0.473 0.737 1.225 1.525

Standard Deviations
37.083 121.563 148.166 187.155 161.976 173.540 185.753 141.604

Crossed-Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1.000</td>
<td>0.881</td>
<td>-0.414</td>
<td>-0.291</td>
<td>0.241</td>
<td>0.820</td>
<td>0.925</td>
<td>0.940</td>
</tr>
<tr>
<td>II</td>
<td>0.881</td>
<td>1.000</td>
<td>-0.569</td>
<td>-0.439</td>
<td>0.049</td>
<td>0.737</td>
<td>0.892</td>
<td>0.943</td>
</tr>
<tr>
<td>V1</td>
<td>-0.414</td>
<td>-0.569</td>
<td>1.000</td>
<td>0.981</td>
<td>0.764</td>
<td>0.022</td>
<td>-0.328</td>
<td>-0.500</td>
</tr>
<tr>
<td>V2</td>
<td>-0.291</td>
<td>-0.439</td>
<td>0.981</td>
<td>1.000</td>
<td>0.843</td>
<td>0.147</td>
<td>-0.205</td>
<td>-0.378</td>
</tr>
<tr>
<td>V3</td>
<td>0.241</td>
<td>0.049</td>
<td>0.764</td>
<td>0.843</td>
<td>1.000</td>
<td>0.639</td>
<td>0.337</td>
<td>0.160</td>
</tr>
<tr>
<td>V4</td>
<td>0.820</td>
<td>0.737</td>
<td>0.022</td>
<td>0.147</td>
<td>0.639</td>
<td>1.000</td>
<td>0.935</td>
<td>0.846</td>
</tr>
<tr>
<td>V5</td>
<td>0.925</td>
<td>0.892</td>
<td>-0.328</td>
<td>-0.205</td>
<td>0.337</td>
<td>0.935</td>
<td>1.000</td>
<td>0.977</td>
</tr>
<tr>
<td>V6</td>
<td>0.940</td>
<td>0.943</td>
<td>-0.500</td>
<td>-0.378</td>
<td>0.160</td>
<td>0.846</td>
<td>0.977</td>
<td>1.000</td>
</tr>
</tbody>
</table>

An eigenvector decomposition of this matrix gave:

Eigenvalues:
4.903 2.876 0.107 0.093 0.009 0.007 0.004 0.002

Fractional Variances:
0.613 0.359 0.013 0.012 0.001 0.001 0.001 0.000

Note, please, that the first two vectors have fractional variance contributions [0.613 and 0.349] that together account for 96.2% of all the variance [or 'information content'] of the data file. It is doubtful in this case whether any of the remaining six vectors contribute further useful information. This is not so, in general, and other examples will present cases with information spread out over 4 or more vectors.

The resulting eigenvectors of these data are, reading downwards along the columns, from left to right:

Eigenvectors:

<table>
<thead>
<tr>
<th>EOF1</th>
<th>EOF2</th>
<th>EOF3</th>
<th>EOF4</th>
<th>EOF5</th>
<th>EOF6</th>
<th>EOF7</th>
<th>EOF8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.431</td>
<td>0.054</td>
<td>0.816</td>
<td>-0.298</td>
<td>-0.171</td>
<td>0.107</td>
<td>0.124</td>
<td>-0.010</td>
</tr>
<tr>
<td>0.432</td>
<td>-0.054</td>
<td>-0.507</td>
<td>-0.714</td>
<td>-0.187</td>
<td>-0.056</td>
<td>-0.068</td>
<td>0.031</td>
</tr>
<tr>
<td>-0.233</td>
<td>0.504</td>
<td>-0.023</td>
<td>-0.012</td>
<td>-0.548</td>
<td>0.452</td>
<td>-0.432</td>
<td>0.011</td>
</tr>
<tr>
<td>-0.177</td>
<td>0.539</td>
<td>-0.088</td>
<td>-0.264</td>
<td>0.384</td>
<td>0.319</td>
<td>0.585</td>
<td>0.095</td>
</tr>
<tr>
<td>0.066</td>
<td>0.582</td>
<td>0.107</td>
<td>-0.074</td>
<td>0.309</td>
<td>-0.620</td>
<td>-0.401</td>
<td>0.022</td>
</tr>
<tr>
<td>0.378</td>
<td>0.312</td>
<td>-0.192</td>
<td>0.389</td>
<td>-0.352</td>
<td>-0.192</td>
<td>0.392</td>
<td>-0.507</td>
</tr>
<tr>
<td>0.439</td>
<td>0.119</td>
<td>-0.118</td>
<td>0.380</td>
<td>-0.054</td>
<td>0.061</td>
<td>0.051</td>
<td>0.791</td>
</tr>
<tr>
<td>0.450</td>
<td>0.009</td>
<td>-0.080</td>
<td>0.175</td>
<td>0.515</td>
<td>0.505</td>
<td>-0.366</td>
<td>-0.327</td>
</tr>
</tbody>
</table>
The elements of these vectors may be thought of as weighting coefficients from which new, orthogonal arrays, EOF(i,j) can be assembled from the observed ECG data as, for the example with the leftmost column, EOF(i,1)

\[
\text{EOF}(i,1) = +0.431 \times \text{ECG}(i,1) + 0.432 \times \text{ECG}(i,2) - 0.233 \times \text{ECG}(i,3) - 0.177 \times \text{ECG}(i,4) \\
... \text{ etc}
\]

\[
\text{ECG}(i,j) = [\text{Obs}(i,j - <\text{Obs}(j)>)] / \text{SD}(j)
\]

\[
\text{Obs}(i,j) = \text{raw observations of the 'raw potential data e1, e2,.. e8.}
\]

\[
<\text{Obs}(j)> = \text{the time-average of Obs}(i,j)
\]

\[
\text{SD}(j) = \text{the standard deviation of Obs}(i,j)
\]

\[
i = 1 \text{ to number of time steps [30,000]}\\n\text{j = 1 to 8 data channels}
\]

These EOF(i,j) are called "Empirical Orthogonal Functions". They may be thought of as transformed ECG potentials Obs(i,j), normalized to zero mean and unit standard deviation. It is a property of EOFs that the integral [or sum over i = 1 to N] of products <EOF(i,j)\cdot EOF(i,k)> are identically zero when j ≠ k. This property reflects the independence of their separate information contents and makes senseless any further linear combinations among different EOFs.
Appendix B: Cluster Analyses

In the present reconnaissance we have 175 separate ECG data records, each of 8 columns [I, II, V1-V6]. It seems unlikely that 175 separately characterizable heart pathologies distinguish them. Can we group the records into a fewer number similar clusters that are separately distinguishable from one another, in useful ways? There are two common algorithms to attempt this.

The first, and more common, is called 'tree' clustering. It begins with finding the pairs of 'objects' [ECG records] that most resemble one another by any of several criteria, then adding other objects in sequence to preceding pairs or triplets, or quartets, [or N-tets], until all hang together from a single apex. Then, working down from the top, one may distinguish 'branches', or families with [presumably] physical similarities. Tree clustering is commonly used, for example, to illustrate evolutionary branching among different species, characterized by DNA similarities and differences.

A second 'K-type' clustering algorithm starts with the presumption that K natural clusters exist. The objects are initially distributed randomly into K piles, then shifted around, one by one, or exchanged in pairs, to maximize some value function. I have processed the 175 ECG 'objects' by this recipe, as follows:

Each object was first characterized by 12 'predictors':

- \( P(1) \) = average heart rate [beats / min]
- \( P(2) \) = standard deviation of heart rate [beats / min]
- \( P(3) \) = maximum of the heart-rate lagged autocorrelation
- \( P(4) \) = skewness of the 1st eigenvector
- \( P(5) \) = kurtosis of the 1st eigenvector
- \( P(6) \) = skewness of the 2nd eigenvector
- \( P(7) \) = kurtosis of the 2nd eigenvector
- \( P(8) \) = a measure of the high-frequency [4 msec] noise
- \( P(9) \) = the fractional variances attributable to the 1st eigenvector
- \( P(10) \) = ditto, the 2nd eigenvector
- \( P(11) \) = ditto, the 3rd
d- \( P(12) \) = ditto, the 4th

Note that these predictors have different dimensions. As we are going to compare 'distances' between objects, we must rescale the predictors in some sensible way. For this
reason all of the \( \mathbf{P}(i) \) vectors have been rescaled to zero mean and unit standard deviation, as in:

\[
\mathbf{p}(i) = [\mathbf{P}(i) - \langle \mathbf{P}(i) \rangle] / \text{Standard Deviation of } \mathbf{P}(i)
\]

\( \langle \mathbf{P}(i) \rangle = (1/N) \sum \mathbf{P}(i) \), \( i = 1 \) to \( N \), the number of observations.

Further, these \( \mathbf{p}(i) \) are not independent of one another. Their crossed-correlation matrix is:

Some of the larger correlations have been \textbf{boldfaced}, for emphasis. The eigenvalues, fractional variances and eigenvectors of this correlation matrix are:

Note that only the 1st five eigenvalues of the \( \mathbf{p}(i) \) exceed [or nearly equal] unity. That is, the 12 original predictors collapse to 5 'significant' combinations of unequally weighted sums and differences of \( \mathbf{p}(i) \). I have emphasized this in the matrix above with \textbf{boldface} type for the larger weights. Note that in the 1st vector the heartbeat rate [Hrate], high-frequency noise [HFn], and the
fractional variances carried by the 1st two data vectors carry most of the weight, and so forth..

Continuing, the 'distance' between any two objects Obj(j) and Obj(k) is defined as

\[ d(k, j) = \frac{1}{12} \left\{ \sum_{i=1}^{12} \left[ p(j, i) - p(k, i) \right]^2 \right\}^{1/2} \]

For every cluster, we define the coordinates of a centroid as \( c(1), c(2), \ldots c(12) \), as

\[ c(j) = \frac{1}{N} \sum_{i=1}^{N} \left[ p(j, i) \right] \]

\[ j = 1 \text{ to } 12 \text{ predictors}, \]
\[ i = 1 \text{ to } N = \text{ the number of objects in that cluster} \]

The distance between the centroids of different clusters \([m, n]\) is

\[ D(m, n) = \left\{ \sum_{j=1}^{12} [c(m, j) - c(n, j)]^2 \right\}^{1/2} \]

\[ \ldots j = 1 \text{ to } 12 \text{ predictors} \]

Finally, an "Separation Index" of the kth cluster SI(k), is defined as

\[ SI(k) = \frac{d(k)}{D(k)} \text{ the rms distance of the k-objects in a cluster and the centroid of that cluster, divided by the distance between that centroid and that of its nearest neighbor.} \]

A K-type cluster algorithm minimizes \( V \), the linear average of the SI(k),

\[ V = \frac{1}{K} \sum_{k=1}^{K} SI(k) \]
\[ k = 1 \text{ to } K, \text{ the number of clusters.} \]

by trial and error, considering sequential transfers of all objects among all the clusters, one at a time, and all pair-wise interchanges. The kth cluster is considered 'significantly separated' to the degree that the Separation Index is less than one:

\[ SI(k) < 1. \]
The present algorithm is distinguished by two additional innovations:

1. As a test of successful clustering, a set of 'faux objects' is invented, using Gaussian-normal predictors, \( p(j) \), and members of this set are initially distributed randomly among the initial clusters. On exit from the analyses, all the faux objects should appear in the least significant cluster, and none should appear in any other. This fairly stringent test was used to detect the largest number, \( K-1 \), of significant clusters, with the last, non-significant, \( K^{th} \) cluster distinguished both by \( S(K) > 1 \) and by its containing all the random faux objects.

2. The algorithm may work the number of objects in any cluster down towards zero. This occurred with the present analysis when the number of initial clusters, \( K \), was over-specified. At \( K = 8 \) the analysis generated two small clusters, each containing a single pair of objects: 22F8882b + 3364CF96 [\( S(1) = 0.213 \)] and 3C4F0F06 + 397DB309 [\( S(2) = 0.373 \)]. I have highlighted these pairings in blue and green, in Table I, section V, of the main text of this note. These may represent duplicate ECG measurements of the same patients. Other duplicates may be present also.