SAND2016-1866C

Trilinear Analysis of Multivariate Electrocardiographic Data

M.H. Van Benthem, R.D. West, T.E. LaBruyere, K.M Simonson, and R.L. Hansen

Sandia National Laboratories, Albuquerque, NM 87185

Abstract

The electrocardiograph (ECG) is an invaluable tool for detecting and diagnosing cardiac irregularities. ECG data are collected using a number of leads, typically 12, attached externally to the chest of the subject or patient. Interpretation of the electrocardiogram (also, ECG) is accomplished predominantly by a detailed examination of the sinus rhythm. In this work, we take a different approach to analyzing ECGs, employing multivariate analysis. Using data from the PhysioNet online database1, we conducted trilinear analysis of 15-lead ECG data. We will present our methods of data preprocessing, method of trilinear analysis and results of data from patients as well as control subjects.

1. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]; 2000 (June 13). PMID: 10851218; doi: 10.1161/01.CIR.101.23.e215

Normal Sinus Rhythm

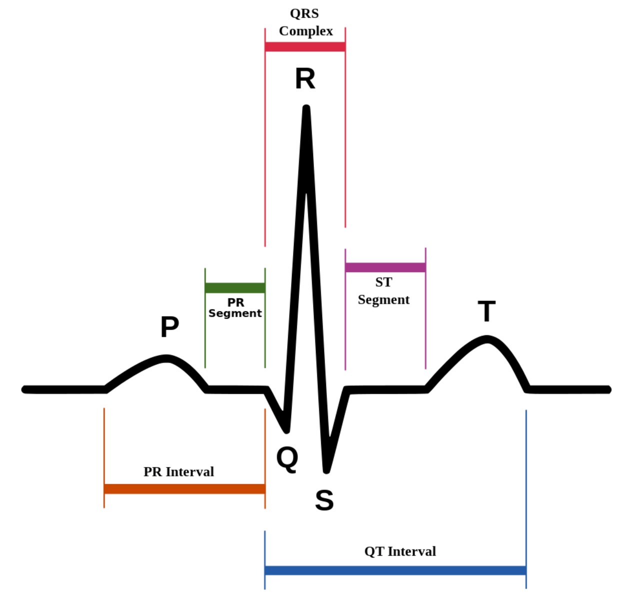
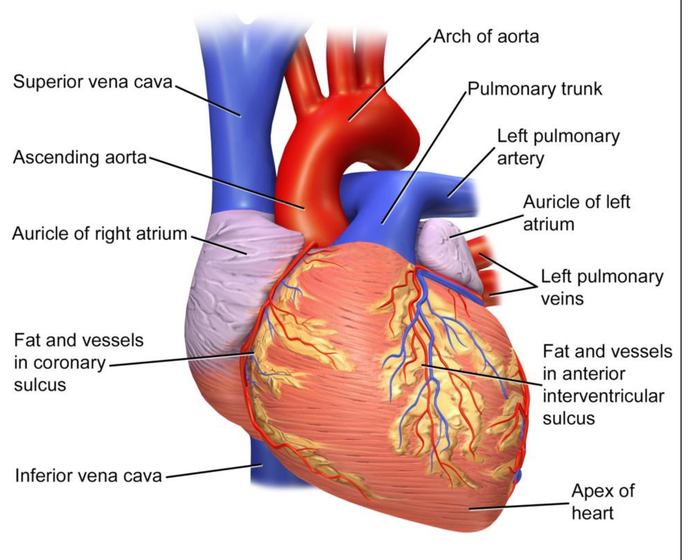


Image source: Created by Agateller (Anthony Atkielski), converted to svg by atom. - en:Image:SinusRhythmLabels.png, Public Domain, https://commons.wikimedia.org/w/index.php?curid=1560893

The Human Heart



Superficial Heart Anatomy (Anterior)

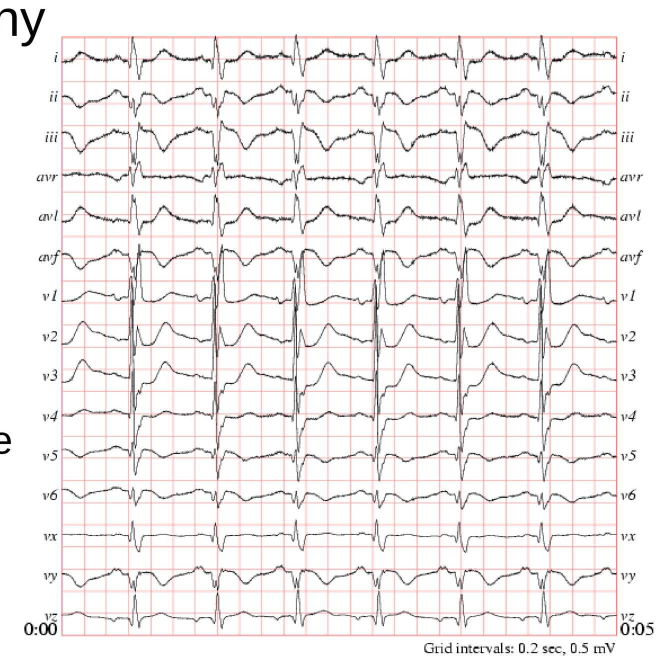
Image source: Blausen Medical Communications, Inc. - Donated via OTRS, see ticket for details, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=26986380

Motivation

- **Explore algorithms that can facilitate interpretation of ECGs**
  - Eliminate bias in ECG data interpretation
  - Use machine learning approach to speed data analysis
- **Develop methods to classify groups using frequentist statistics**
  - Impose rigorous statistical approach to avoid misclassification
  - Use one-class classifier rather than Bayesian methods
- **Employ factor analysis techniques on challenging datasets**
  - Expand the realm of data prospects open to multivariate analysis
  - Combine multivariate and multiway analysis with new classification algorithms

ECG Data

- **Physikalisch-Technische Bundesanstalt (PTB)**
  - National Metrology Institute of Germany
  - Compilation of digitized ECGs
  - Database
    - 549 records from 290 subjects
      - Aged 17 to 87, mean 57.2
      - 209 men, mean age 55.5
      - 81 women, mean age 61.6
      - Ages not recorded for 1 female and 14 male subjects
    - Each record includes 15 simultaneously measured signals
      - Conventional 12 leads
      - 3 Frank leads



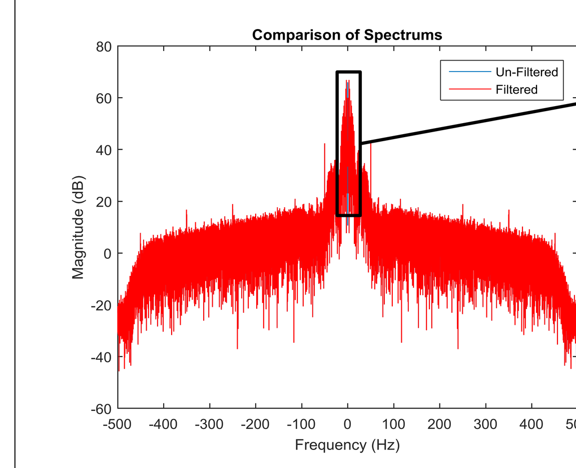
http://www.physionet.org/physiobank/database/ptbdb/

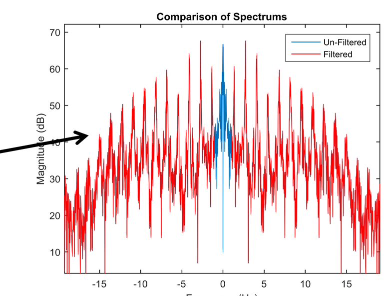
Data Processing

ECG Baseline Correction and Noise Filtering

- The ECG data have low-frequency trends
- Through the use of meta data and some experimenting, it appears that low-frequency trends are below ~0.5 Hz
- Signal processing methods are used to remove the low-frequency trends and stabilize the ECG data
- The proposed method seems to be robust and reliable and has little impact on the ECG signal

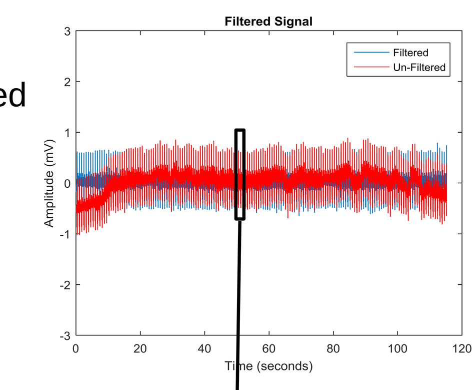
Filtering in Fourier Domain



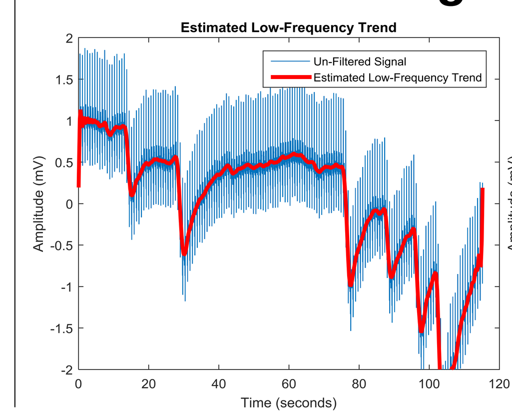


- The low-frequency components are removed by zeroing-out FFT bins
- Note that the signal is highly oversampled

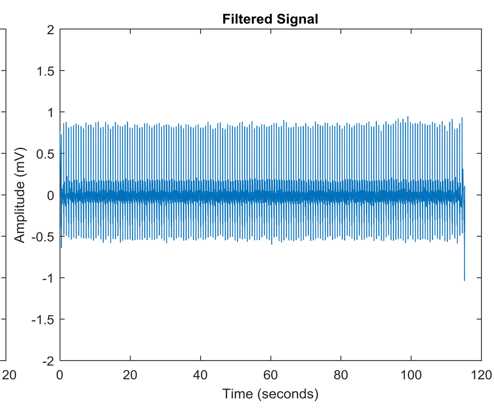
- Note: The signal is being filtered, too. However, the features seem to be high frequency, thus they are preserved.

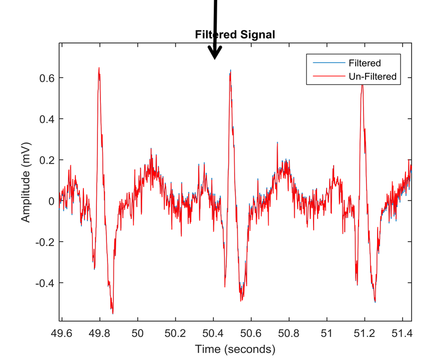


Before Filtering



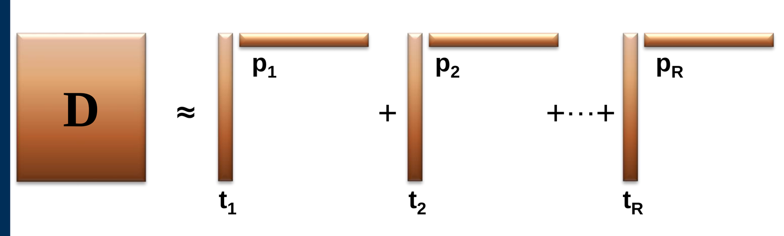
After Filtering





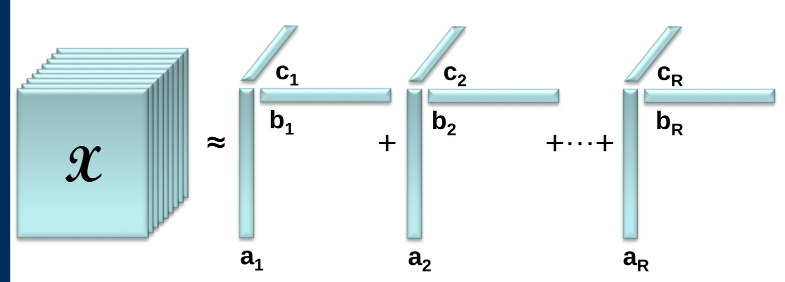
Data Analysis

Two-Way Analysis Methods[1]



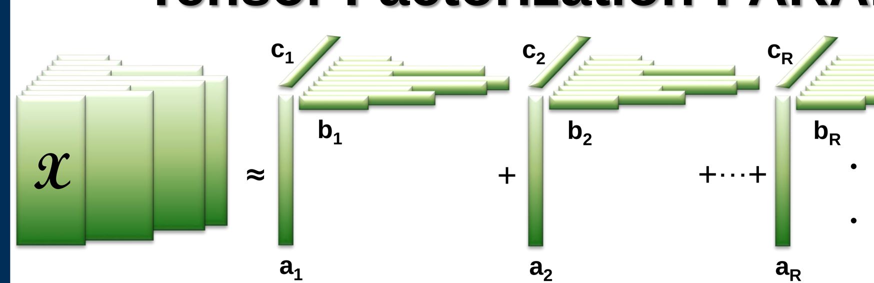
- Principal Component Analysis (PCA)
  - Given a matrix containing data, D, as a first step in many analyses we want principal components
    $$\mathbf{D} \cong \mathbf{T}\mathbf{P}^{\mathrm{T}}$$
  - Such that T and P are an orthogonal basis sets, that is a reduced dimensional representation of D, with ordered maximized variance.
    - T is orthogonal (scores); P is orthonormal (loadings).
- Multivariate Curve Resolution (MCR)
  - Impose constraints on solution space

Tensor Factorization-PARAFAC[2]



- Tensor factorizations of multi-way data
  - Parallel factor analysis (PARAFAC)
  - Nonnegative tensor factorization (NTF)
  - Similar in idea to least squares matrix techniques: principal component analysis (PCA), singular value decomposition (SVD), multivariate curve resolution (MCR)
- When applied to a data array, data are modeled as a mixture of factors, each with its own triad of signature factors

Tensor Factorization-PARAFAC2[3]



- PARAFAC2 permits factorization of unaligned or different length signals
- This data appears to not follow a model suitable for PARAFAC
  - Temporal alignment of sinus rhythms not trivial
  - Lack of alignment may prevent use of standard PARAFAC algorithm

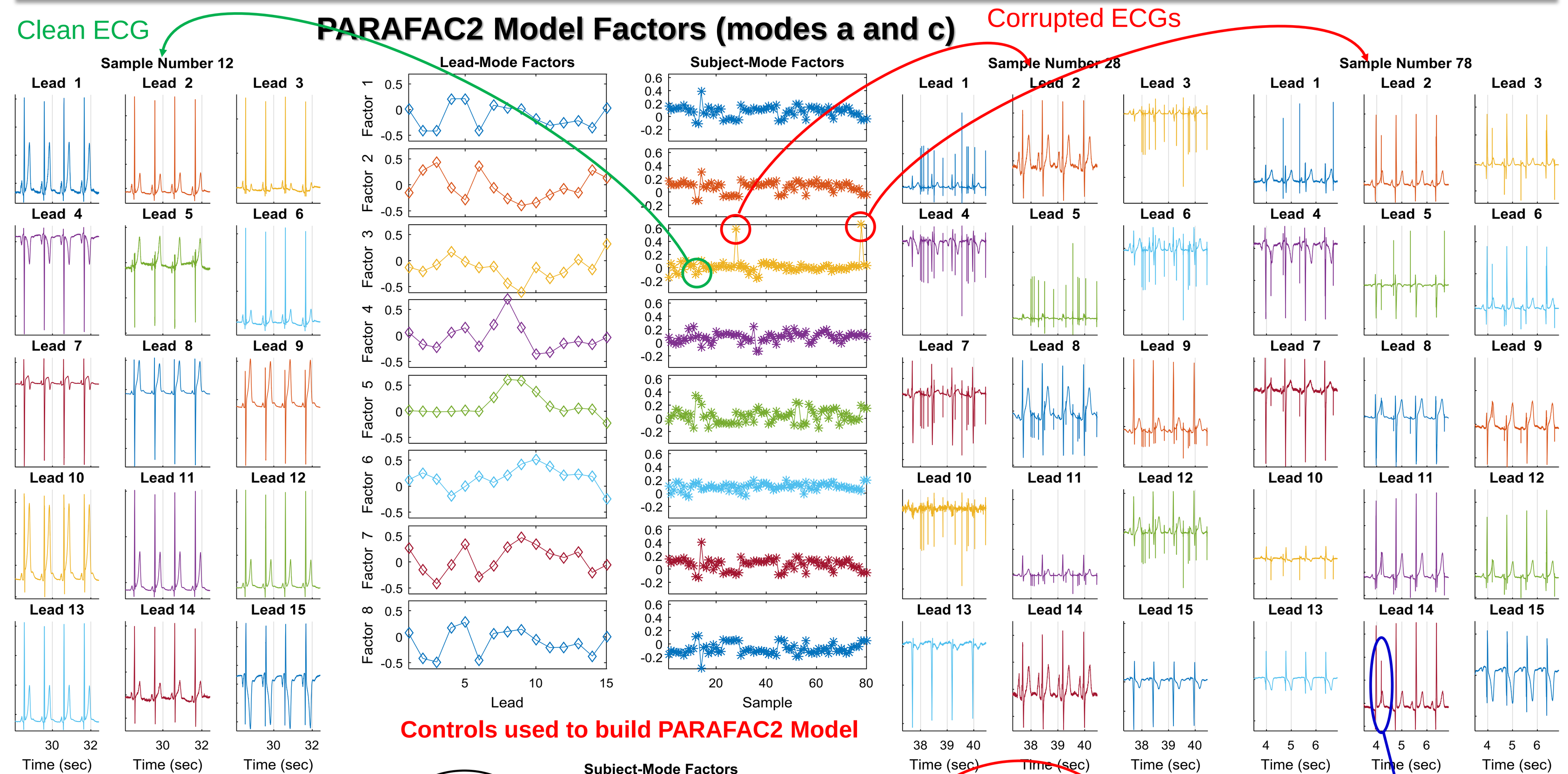
PARAFAC2 analysis conducted with a SNL-developed algorithm

- **ECG Data Utilized for PARAFAC2**
  - 494 records from 244 subjects
    - Records with minimum 90 seconds (0.001 sec increments)
    - Used 80 seconds of each record after first 4 seconds
    - Array size 15 x 80,000 x 494
  - 80 records from 52 Controls
    - Construct PARAFAC2 model using Controls
    - Model array size 15 x 80,000 x 80
  - 414 records from 192 Patients
    - Predict patient factors from Control-based model

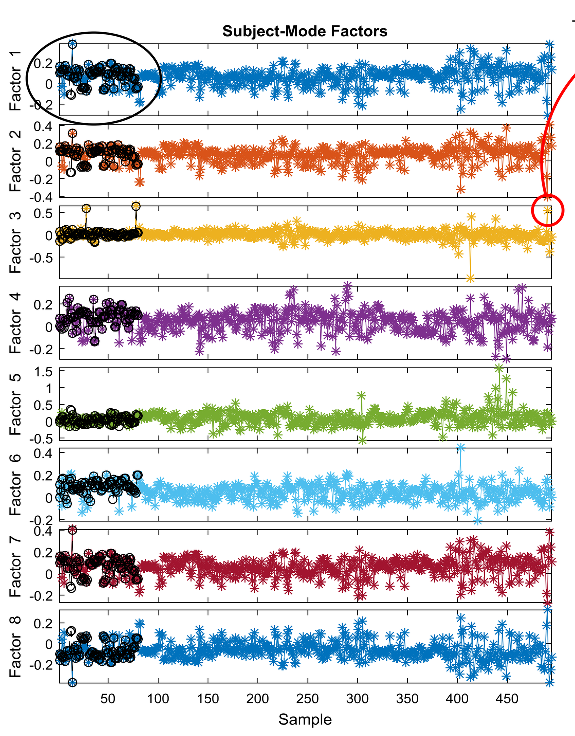
1. M. R. Keenan, [Multivariate Analysis of Spectral Images Composed of Count Data] John Wiley & Sons, Ltd, Chichester, West Sussex, England(2007).
2. T. Kolda, and B. Bader, "Tensor Decompositions and Applications," SIAM Review, 51(3), 455-500 (2009).
3. H. A. L. Kiers, J. M. F. TenBerge, and R. Bro, "PARAFAC2 - Part I. A direct fitting algorithm for the PARAFAC2 model," J. Chemom., 13(3-4), 275-294 (1999).

Results

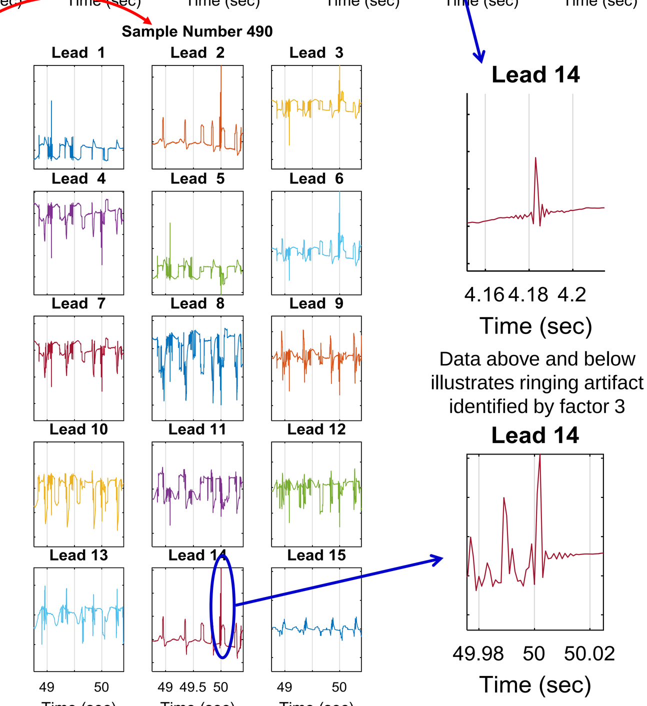
Clean ECG — PARAFAC2 Model Factors (modes a and c) — Corrupted ECGs



Controls used to build PARAFAC2 Model



- Factor 3 appears to strongly correlate with an electrical artifact with leads or ECG electronic "ringing".
  - Both ECGs with high values in Factor 3 show artifact, others do not exhibit this behavior
- PARAFAC2 model-derived factors in black
  - PARAFAC2 model built with control subjects
  - Note that re-estimated values have some ambiguity
  - Most values are identical to model
- Colored asterisks are the estimated values using PARAFAC2 projection
  - Samples with index higher than 80 are patients with some heart disease
- Identification of any specific disease condition is not readily identifiable
- General "healthy" versus "disease" classification is not clear



Data above and below illustrates ringing artifact identified by factor 3

Conclusions

- **Processed and performed multivariate analysis on multi-lead ECG data**
  - Developed Matlab®-based pseudo-code to import PhysioBank ECG data
    - Imports data significantly faster than available Java-based tools
  - Developed FFT-based approach to baseline correction and noise removal
    - Eliminates irregular baseline and low-frequency noise
- **Developed fast algorithm for performing PARAFAC2 for large datasets**
  - Employs core-PARAFAC tactic - speeds up calculations
- **Performed PARAFAC2 modeling of 15-lead ECG data**
  - Readily identified data artifacts in ECGs
  - Attempts to combine multivariate and multiway analysis with one-class classifier algorithms not successful
    - No clear delineation between control subjects and subjects with heart disease revealed in analysis.

[MATLAB] The MathWorks, Inc., Natick, MA(2015).

Acknowledgements

The Sandia National Laboratories Laboratory Directed Research and Development (LDRD) program provided funding for this project. Sandia is a multiprogram laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Contact mark.vanbenthem@sandia.gov for reprints, questions, or more information