Smoothing and Discriminating MAP Data

K Jin, N Stockbridge

FDA/CDER, White Oak, MD, USA

Abstract

Cubic B-Splines are used to approximate mean ambulatory blood pressure data and the fitted coefficients serve as discretization of the curves. A rank based method is proposed to predict Test Set A. A double cross validation approach is proposed to predict Test Set B.

1. Introduction

The goal of Computers in Cardiology Challenge 2009 [3] is to predict acute hypotensive episodes (AHE), which are defined as any period of 30 or more minutes during which at least 90% of mean arterial blood pressure measurements are ≤60 mmHg. The Challenge provides three datasets: a Training Set, Test Set A and Test Set B. The Training Set has data collected from 60 patients, of whom 30 patients, denoted as "H", developed AHE during the forecast window. The Test Set A has data from 10 patients, and one is supposed to predict 5 "H" from this set. Test Set B has 40 patents, and one is supposed to predict 10 to 16 "H" from this set.

All datasets include continuous telemetric data for heart rate, systolic and diastolic pressure, and mean arterial blood pressure (MAP), and other clinical information on vital signs, concomitant medications, etc. We used only the mean blood pressure data.

In this paper, we propose a two-stage approach to this challenge. The first step is to approximate the mean arterial blood pressure curve by cubic B-Splines. The resulting coefficients corresponding to the bases are considered to be a discretization of the continuous MAP. The second step is to find the "best subset" of the discretization from the Training Set that best discriminated "H" from "C". This "best subset" is then used to predict the "H" from the two test sets.

2. Cubic B-Splines and discretization

Cubic B-Splines[1] are used to approximate MAP curves. B-Splines generally reflect the local features of the target curve.

Technically, the t-axis of original ABP mean curves are inverted so that the t = 0 is the T_0, the starting point of the prediction window.

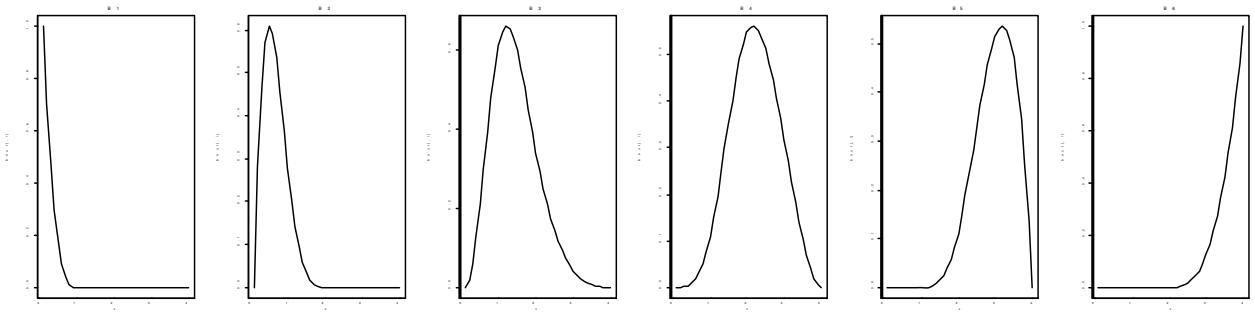

Figure 1. This figure show cubic B-spline bases with equal spaced k=3 knots.

Let $B_k(t), k = 1, ..., K+3$ be bases with equally spaced K knots. Denote $ABPM_i(t_j), j = 1, ...n_i$ to be i^{th} MAP curve. $\sum_{k=1}^{K} \alpha_{ik} B_k(t_j)$ is used to approximate the MAP curve and α_{ik} is a least square estimate by minimizing $\sum_{j=1}^{n_i} (\sum_{k=1}^{K} \alpha_{ik} B_k(t_j) - ABPM_i(t_j))^2$.

To select the smoothing parameter, a generalized cross validation criterion $\hat{GCV}_i(\mathrm{K}) = \hat{RSS}/(n + 1 - 2.5 * (K-1))$ is calculated for each $i, i = 1, ..., N$. Here N is 70 to predict Test Set A by adding Test Set A to the Training Set, and N is 100 to predict Test Set B. The smoothing

parameter K is selected by minimizing $\sum_{i=1}^{N} \hat{GCV}_i$ (K). Because of the different lengths of the training datasets, the smoothing parameter selection is done first on the minimal common interval of length of T=656. The remaining intervals are then assigned equally spaced knots generated from the common interval

The Training Set, Test Set A and Test Set B are fitted with the selected K, the result of (α_{ik}) matrices are denoted as $(\alpha)^T_{60,K}$, $(\alpha)^A_{10,K}$ and $(\alpha)^B_{40,K}$, respectively.

These matrices are considered to be discretizations of MAP curves and will preserve the features of these curves that can be used to carry out discrimination analysis.

3. Rank based discrimination

A simple rank based discrimination was developed from the Training Set and used to predict Test Set A. We first select those columns in $(\alpha)^T_{60,K}$ that have high discriminating power, to distinguish H and C cases. We rank each column separately and then look at the distribution of H and C. Those columns with high concentration of H or C at the top or bottom are considered to have high discrimination power.

Let $I = (i_1, ..., i_{60})$,where $i_j = -1, (j = 1, ..., 30), 1$ otherwise; $\vec{\alpha}_k = (\alpha_{1k}, ..., \alpha_{60k})$ be the k^{th} column of $(\alpha)^T_{60,K}$. The order statistics of $\vec{\alpha}_k$, where $i(k)$ is the original location of $\alpha_{i(k)k}$ in $\vec{\alpha}_k$.

Then $I_{(k)} = (i_{1(k)}, ..., i_{60(k)})$ is a transformation of $\vec{\alpha}_{(k)}$, where entry corresponding to H becomes -1 and entry corresponding to C becomes 1. Large absolute values of $|\sum_{j=1}^{30} i_{j(k)}|$ and $|\sum_{j=31}^{60} i_{j(k)}|$ result from a large concentration of similar group at the top or bottom. We define $\Delta_k = |\sum_{j=1}^{30} i_{j(k)} - \sum_{j=31}^{60} i_{j(k)}|$, so that a large value of Δ_k indicates high discrimination power of $\vec{\alpha}_k$.

3.1. Application to Test Set A

First, we find which columns have the highest discrimination power from the training matrix $(\alpha)^T_{60,K}$. Then, Δ_k is calculated for all columns of $(\alpha)^T_{60,K}$ and the result is shown in Table 1.

0 8 20 16 24 16 27 11 27 12 24 24 16 24 20 8 8 4 20 7 16 8 12 16 24 16 12 16 28 16 12 20 20 24 28 16 16 24 16
24 16 20 16 24 16 20 12 16 15 18 22 10 21 13 13 13 5 21 13 16 15 11 23 19 23 19 22 18 18 22 10 18 17 16 12 16 12
4 8 11 14 10 16 12 4 7 15 9 5 12 8 15 14 9 5 9 1 9 5 7 15 15 3 15 3 15 6 26 5 6 8 4 4 0 4 4 8 4 12 3 9
13 2 10 6 10 6 17 1 9 9 2 6 2 14 3 4 4 4 2 9 5 1 1 5 3 3 1 0 3 3 7 7 11 3 3 3 7 3 3 3 1 2 2 3 5 5
3 3 3 3 3 7 7 1 2 2 2 2 2 2 6 6 6 3 1 3 5 1 3 1 7 3 3 10 6 3 1 3 3 3 1 5 5 5 1 8 5 12 5 7 5 7 1
7 3 7 1 0 1 7 1 2 2 6 3 3 1 3 2 6 6 6 2 2 1 3 1 5 5 5 5 1 5 1 7 1 1 5 1 ....

Table 1. Δ_k from $(\alpha)^T_{60,K}$

The 15 columns that have the largest Δ_k's are columns 5,7,9,11,12,14,25,29,34,35,38,40,44,63,65. These columns from matrix $(\alpha)^A_{10,K}$ are selected and ordered separately. The results from application to Test Set A are displayed in Table 2. Except for column 29, every ordered column puts cases 101, 102, 104, 109 and 110 at the top portion that predicts them as "H". On column 29, cases 101, 102, 108, 109, 110 are predicted as "H". From these results, cases 101, 102, 104, 109 and 110 are predicted as "H" and these answers are correct.

K=5	7	9	11	12	14	25	29	34	35	38	40	44	63	65
110	110	110	110	102	110	101	101	101	101	101	101	101	102	104
109	101	101	102	109	101	109	110	109	110	102	109	109	104	101
104	104	109	101	110	109	110	109	110	109	109	110	104	101	110
101	109	104	109	104	102	104	102	102	104	110	102	110	109	109
102	102	102	104	101	104	102	108	104	102	104	104	102	110	102
108	108	108	108	108	108	108	104	108	108	108	108	108	108	108
107	107	107	107	107	107	107	107	107	107	107	107	107	107	107
105	105	105	105	105	105	105	105	105	105	105	105	105	105	105
103	103	103	103	103	103	103	103	103	103	103	103	103	103	103
106	106	106	106	106	106	106	106	106	106	106	106	106	106	106

Table 2. Ordering the corresponding column in Test Set A.

4. Logistic regression and cross-validation

For Test Set B, the ranking approach did not generate a conclusive result. Therefore, a different approach was proposed. By assigning 1 to "H" and 0 to "C", the status could be naturally linked to $(\alpha)^T_{60,K}$ with logistic regression. The problem becomes how to search these columns from $(\alpha)^T_{60,K}$ that could be used to predict the status. The optimal property should not be judged by its own status but by the statuses of the other data sets. With this setup, the problem falls into a classical "best subsets" selection problem. There are many potential different solutions to apply to it. Here, we used a double cross-validation approach.

4.1. Leave a row out

First, we will define criteria that could be used to select appropriate columns.

Denote Y and $(\alpha)^T_{60,K} = (\alpha_1,...,\alpha_{60})^t$, where α_k is a vector of coefficients of k^{th} MAP curve. For each i, remove y_i from Y, α_i from $(\alpha)_{60,K}$. Fit the remainder Y_{-i} to $(\alpha)^t_{-i}$ with logistic regression, the coefficient is denoted as β_{-i}. Then, y_i can be predicted by $y_{-i} = \exp^{\beta_{-i}\alpha_i} / (1+\exp^{\beta_{-i}\alpha_i})$.

Repeat this procedure for all $i, i = 1,...,60$ in the Training Set, so that we obtain the prediction of Y by $Y_p = (y_{-1},..., y_{-60})^t$.

We now define $CVF_c(\alpha) = mean(|Y - Y_p|)$ and $CVF_d(\alpha) = \#of\ y_{-i}$ such that $|y_i - y_{-i}| \le 0.5$. $CVF_c(\alpha)$ is a continuous measure for closeness of Y_p to Y, while the discrete measure of $CVF_d(\alpha)$ counts how many correct predictions are made if we assign 1 if y_{-i} >0.5 and 0 otherwise.

These two statistics will be used to guide the selection of columns from $(\alpha)^T_{60,K}$.

4.2. Leave a column out

Starting with full columns of $(\alpha)^T_{60,K}$, we will delete less optimized columns repeatedly until no further improvement can be made.

Denote $(\alpha)^T_{60,K} = (\alpha_1,...,\alpha_K)$, where α_k is k^{th} column. For each $k = 1,...,K$, remove α_k from $(\alpha)^T_{60,K}$. $CVF_c(\alpha_{-k})$ and $CVF_d(\alpha_{-k})$ were then calculated, where α_{-k} is the remainder matrix of $(\alpha)^T_{60,K}$ kth column removed. From **Session 4.1**, the smallest $CVF_c(\alpha_{-k})$ or largest $CVF_d(\alpha_{-k})$ indicates that removing the corresponding column will produce a better prediction.

We start with full columns and delete a column each time with one of the following strategies: 1) remove the column with the smallest $CVF_c(\alpha_{-k})$; 2) remove the column with the largest $CVF_d(\alpha_{-40}) = 43$. For strategy 2, when there are tiers, we remove all the columns at the early stages. When tiers occur at the late stages, we remove the column with the smallest $CVF_d(\alpha_{-k})$. The procedure is continued until no further improvement can be made. The remaining columns are considered the best prediction columns to denote $(\alpha)_{60,K-l}$.

We do logistic regression of Y on $(\alpha)_{60,K-l}$ to get coefficient β_{K-l}. The y^B is predicted by $e^{(\alpha)^B_{60,K-l}\beta_{K-l}} / (1+e^{(\alpha)^B_{60,K-l}\beta_{K-l}})$.

4.3. Application to Test Set B

The first curve was removed from the Training Set because of a large missing segment near the beginning. Thus, 59 curves were used in the cross-validation process.

The initial leave-a-row-out procedure returns the largest $CVF_d(\alpha_{-40}) = 48$. After removing column 40, the second run returns the largest $CVF_d(\alpha_{-26}) = CVF_d(\alpha_{-33}) = CVF_d(\alpha_{-44}) = 55$. After removing columns 26, 33 and 44, the procedure returns $CVF_d(\alpha_{-22}) = 58$. After removing column 22, the procedure did not provide further improvement.

Denote $s = (22, 26, 33, 40, 44)$. Let β_{-s} be the coefficient of logistic regression of Y on $(\alpha)_{60,K-s}$. The y^B is predicted by $e^{(\alpha)^B_{60,K-s}\beta_{-s}} / (1+e^{(\alpha)^B_{60,K-s}\beta_{-s}})$.

Simply predicting "H" by $|y_i^B - 1) < 0.5$ seemed to overestimate the number of "H". The result is refined by the ranking method. For "H" that y_i^B is close to 0.5, only these cases that are also predicted as "H" by the ranking method are assigned as "H". The adjustments generate three entries containing 11, 13 and 13 "H" cases. All three entries predict 33 out of 40 correctly.

5. General discussions

This manuscript proposes a two-stage method for predicting AHE from continuous MAP data. In statistical literature, most two-stage approaches can eventually be improved by joining two steps into a simultaneous process. The two-stage approach proposed here is different from these types of approaches. B-Splines fitting is used to find the best fit for the curve, and the second stage finds the best predictors from the fitted coefficients. The best predictors may not necessarily be the best base to fit the curve.

The proposed approach has the potential to be employed in searching for hidden clinical features that may not be easily seen from a digitized wave dataset. Indeed, our approach does not assume any prior knowledge about the dataset or its clinical interpretation.

It may be possible to do much better in predicting clinical events if we utilize other channels—heart rate, systolic and diastolic pressure, or other clinical telemetry data—but we have not explored how much these add, or even if the MAP data were the optimal single channel.

In addition, the proposed methods could be further improved by careful consideration of how to access variances in the data sets and estimates.

The training datasets have various lengths, from 656 to 14,986. The logistic regression and cross-validation methods we used require each dataset to be the same

length. Therefore, only limited data segments with the common length of 656 are used in the logistic regression and cross-validation. This limitation can probably be overcome.

Interestingly, we note that in the results of both sets A and B, knots at early times were mostly selected for prediction. This suggests some power for predicting AHE further temporally removed.

Disclaimer

The views expressed in this manuscript are those of the authors and not necessarily those of the Food and Drug Administration.

References

[1] Carl de Boor. A Practical Guide to Splines. Springer. 2001.

[2] Jin, K. Empirical smoothing parameter selection in adaptive regression. Ann. Statist. 1992;20:1844-1874

[3] Moody GB, Lehman LH. Predicting acute hypotensive episodes: the 10th annual PhysioNet/Computers in Cardiology Challenge. Computers in Cardiology 2009;36.

Address for correspondence
Kun Jin, Ph.D.
FDA, 10903 New Hampshire Ave., Bldg 21, RM 4622, Silver Spring, MD 20993-0002
kun.jin@fda.hhs.gov