

Hamparsum Bozdogan

Statistical Data Mining and Knowledge Discovery

CRC PRESS

Boca Raton Ann Arbor London Tokyo



Contributors

Slavkovic



Contents

1 Automated Scoring of Polygraph Data	1
<i>Aleksandra B. Slavkovic</i> Department of Statistics, Carnegie Mellon University, Pittsburgh, USA	
1.1 Introduction	1
1.2 Background	2
1.2.1 The Polygraph Examination	2
1.2.2 Instrumentation and Measurements	3
1.2.3 Chart Evaluations	4
1.3 Statistical Models for Classification and Prediction	5
1.4 The Data	7
1.5 Statistical Analysis	10
1.5.1 Signal Processing	10
1.5.2 A Simplified Approach to Feature Extraction	11
1.5.3 Feature evaluation, modeling, and classification	13
1.5.4 Logistic Regression	14
1.5.5 Classification Results	15
1.6 Discussion	16
1.7 Conclusion	18



1

Automated Scoring of Polygraph Data

Aleksandra B. Slavkovic

Department of Statistics, Carnegie Mellon University, Pittsburgh, USA

CONTENTS

1.1	Introduction	1
1.2	Background	2
1.3	Statistical Models for Classification and Prediction	5
1.4	The Data	6
1.5	Statistical Analysis	10
1.6	Discussion	16
1.7	Conclusion	18
	Acknowledgments	19
	References	19

The objective of automated scoring algorithms for polygraph data is to create reliable and statistically valid classification schemes minimizing both false positive and false negative rates. With increasing computing power and well developed statistical methods for classification we often launch analyses without much consideration for the quality of the datasets and the underlying assumptions of the data collection. In this paper we try to assess the validity of logistic regression when faced with a highly variable but small dataset of 149 real-life specific incident polygraph cases. The data exhibit enormous variability in the subject of investigation, format, structure, and administration, making them hard to standardize within an individual and across individuals. This makes it difficult to develop generalizable statistical procedures. We outline steps and detailed decisions required for the conversion of continuous polygraph readings into a set of features. With a relatively simple approach we obtain accuracy rates comparable to those reported by other more complex algorithms and manual scoring. Complexity underlying assessment and classification of examinee's deceptiveness is evident in a number of models that account for different predictors giving similar results, typically "overfitting" with the increasing number of features. While computerized systems have the potential to reduce examiner variability and bias, the evidence that they have achieved this potential is meager at best.

1.1 Introduction

Polygraphs are used by law enforcement agencies and the legal community for criminal investigations, in the private sector for pre-employment screening, and for testing

for espionage and sabotage. Polygraph proponents claim high accuracy rates of 98% for guilty subjects and 82% for innocent [23, 2]. These rates are typically calculated by leaving out inconclusive cases and ignoring the issue of sampling bias, e.g. when the accuracies and inter-raters reliability are calculated using only subjects for which there is an independent validation of their guilt or innocence (i.e., ground truth).

The polygraph as an instrument has been recording changes in people's relative blood pressure, respiration and the electrodermal response (palmar sweating) in some form since 1926. These psychophysiological responses, believed to be controlled by the autonomic nervous system, are still the main source of information from which the polygraph examiners deduce an examinee's deceptive or non-deceptive status. The underlying premise is that an examinee will involuntarily exhibit fight-or-flight reactions in response to the asked questions. The autonomic nervous system will, in most cases, increase the person's blood pressure and sweating, and affect the breathing rate. These physiological data are evaluated by the polygraph examiner using a specified numerical scoring system and/or statistically automated scoring algorithms. The latter are the main focus of this report. Current methods of psychophysiological detection of deception (PDD) are based on years of empirical work, and are often criticized for a lack of thorough scientific inquiry and methodology.

The objective of automated scoring algorithms for polygraph data is to create reliable and statistically valid classification schemes minimizing both false positive and false negative rates. The statistical methods used in classification models are well developed, but to the author's knowledge, their validity in the polygraph context has not been established. We briefly describe the polygraph examination framework and two automated scoring algorithms relying on different statistical procedures in Section 1.2. Section 1.3 provides some background on statistical models one might naturally use in settings such as automated polygraph scoring. In Section 1.4 we evaluate collection of real-life polygraph data of known deceptive and nondeceptive subjects. In Section 1.5 we outline steps and detailed decisions required for the conversion of continuous polygraph readings into a set of numeric predictor variables and present results of a logistic regression classifier. Our approach is simpler than other proposed methods, but appears to yield similar results. Various data issues that are not addressed or captured by the current algorithms indicate a deficiency in validity of methods applied in the polygraph setting.

1.2 Background

1.2.1 The Polygraph Examination

The polygraph examination has three parts [24, 20, 6]. The *pre-test* phase is used to formulate a series of 8 to 12 "Yes and No" questions which are custom-made for the examinee based on the exam topic (see Table 1.1). During the *in-test* phase, these questions are asked and data are collected. Typically, the same questions are asked

at least 3 times, with or without varying the order of the questions. One repetition represents a polygraph chart limited to approximately 5 minutes (Figure 1.1).

There are three main types of questions. *Irrelevant* (“Is today Tuesday?”) questions are meant to stabilize the person’s responses with respect to external stimuli such as the examiner’s voice. *Relevant* (R) questions address the main focus of the exam. *Comparison* (C) or *control* questions address issues similar in nature but unrelated to the main focus of the exam. There might be “wild-type” questions too [6].

TABLE 1.1

Example questions for the Zone Comparison Test (ZCT) format.

X	This test is about to begin.
1	Is today Tuesday?
2	Regarding that stolen property, do you intend to answer truthfully each question... ?
3	Are you convinced that I will not ask you a surprise question on this test?
4C	Before 1995, did you ever steal anything from an employer?
5R	Did you steal any of that missing property?
6C	Before 1995, did you ever steal anything?
7R	Did you steal any of that missing property from building — ?
1A	Is this the month of January?
8C	Before 1995, did you ever steal something and not get caught?
9R	Do you know for sure who stole any of that missing property?
10C	Before 1995, did you ever tell a serious lie to ...?
11	Is there something else you are afraid I will ask you a question about...?
XX	Test is over...

In the *post-test* part the examiner evaluates the charts by comparing the responses on the relevant and comparison questions, and tries to obtain a confession from the examinee. It is expected that a deceptive person will show stronger reactions to the relevant questions, while an innocent person will be more concerned with comparison questions. Depending on the test format, the agency conducting the exam and examiner’s training, the number, type and order of questions may differ.

Specific issue exams address known events that have occurred, e.g. a theft. We are concerned with two Comparison Question Test formats: the Zone Comparison Test (ZCT), and the Multiple General Question Test (MGQT). According to the Department of Defense Polygraph Institute (DoDPI), these have pre-defined formats, although in practice they are highly variable (cf. § 1.4).

1.2.2 Instrumentation and Measurements

A polygraph instrument records and filters the original analog signal. The output is a digital signal, a discretized time series with possibly varying sampling rates across instruments and channels. The polygraph typically records thoracic and abdominal respirations, electrodermal and cardiovascular signals (Figure 1.1).

Pneumographs positioned around the chest and the abdomen measure the rate and depth of respiration. Subjects can control their breathing and influence the recorded measurements. Changes in respiration can also affect heart rate and electrodermal activity. For example coughing is manifested in electrodermal activity.

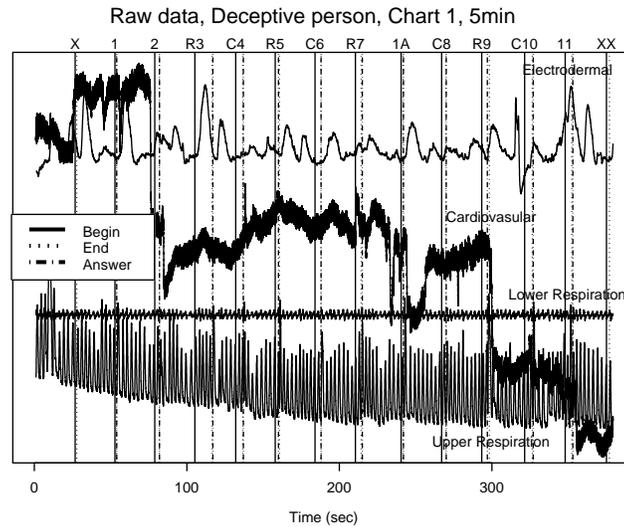


FIGURE 1.1

The lower recordings are thoracic and abdominal respirations, the middle series is cardiovascular signal and the upper is electrodermal signal. The scale is arbitrary. The labels on the upper axis correspond to the Table 1 question sequence.

Electrodermal activity (EDR) or sweating is measured via electrodes (metal plates) placed on two fingers, and it is considered the most valuable measure in lie detection. Either skin conductance (SC) or its reciprocal, skin resistance (SR) is recorded. Some have argued [8] that the size of the response on a question depends on which of SC or SR is recorded. This is a controversial issue discussed in more detail in the psychophysiological literature and in [6].

Cardiovascular activity is measured by a blood pressure cuff placed above the biceps. As a hybrid signal of relative blood pressure and heart rate, it is the most complex of the four measurements. The cardiovascular response is coactivated or coinhibited by other physiological responses, making its evaluation more difficult, e.g. [0.12Hz-0.4Hz] frequency band in the heart rate is due to respiration. This coupling may differ within a person and across different environmental settings [3].

It is unclear whether these physiological responses reflect a single psychological process (such as arousal) or the extent to which they are consistent across individuals. The psychophysiological literature includes contradictory claims on how internal emotional states are mapped to physiological states, and the extent to which emotional states represent deception [6, 11, 17, 14].

1.2.3 Chart Evaluations

A critical part of polygraph examination is the analysis and interpretation of the physiological data recorded on polygraph charts. Polygraph examiners rely on their

subjective global evaluation of the charts, numerical methods and/or computerized algorithms for chart scoring.

Numerical Scoring. The scoring procedure may differ by the PDD exam type, the agency, and the examiner's training and experience. In the 7-Position Numerical Analysis Scale the examiner assigns values from -3 to 3 to the differential responses on pairs of relevant and comparison questions. The negative values, for example, indicate higher reaction on the relevant questions. The values are summed across pairs, channels and charts. A total score of +6 or greater indicates nondeception, -6 or less deception, and in-between are inconclusive cases [20, 25].

Computerized Scoring Algorithms. We focus on two computerized polygraph systems currently used with U.S. distributed polygraph equipment. Other systems have been developed more recently [24, 9]. The Stoelting polygraph instrument uses the Computerized Polygraph System (CPS) developed by Scientific Assessment Technologies based on research conducted at the University of Utah [18, 19, 5]. The Axciton and Lafayette instruments use the PolyScore algorithms developed at the Johns Hopkins University Applied Physics Laboratory [12, 13, 22]. Performance of these algorithms on an independent set of 97 selected confirmed criminal cases was compared by [9]. CPS performed equally well on detection of innocent and guilty subjects while the other algorithms were better at detecting deceptives (cf. § 1.6). More details on the polygraph instruments and history of the development of computerized algorithms can be found in [20, 19, 1].

The methods used to develop the two existing computer-based scoring algorithms both fit within the general statistical framework described below. They take the digitized polygraph signals and output estimated probabilities of deception. While PolyScore uses logistic regression or neural networks to estimate these probabilities, CPS uses standard discriminant analysis and a naive Bayesian probability calculation (a proper Bayesian calculation would be more elaborate and might produce markedly different results). They both assume equal a priori probabilities of being truthful and deceptive. The biggest differences that we can discern between them are the data they use as input, their approaches to feature development and selection, and the efforts that they have made at model validation and assessment. Appendix G of [6] and [24] give a more detailed review of these algorithms. Computerized systems have the potential to reduce bias in the reading of charts and inter-rater variability. Whether they can actually improve accuracy also depends on how one views the appropriateness of using other knowledge available to examiners, such as demographic information, historical background of the subject, and behavioral observations.

1.3 Statistical Models for Classification and Prediction

This section provides some background on the statistical models that one might naturally use in settings such as automated polygraph scoring. The statistical methods

for classification and prediction most often involve structure:

$$\text{response variable} = g(\text{predictor variables, parameters, random noise}), \quad (1.1)$$

where g is some function. For classification problems it is customary to represent the response as an indicator variable, y , such that $y = 1$ if a subject is deceptive, and $y = 0$ if the subject is not. Typically we estimate y conditional on the predictor variables, X , and the functional form, g . For linear logistic regression models, with k predictor variables $x = (x_1, x_2, \dots, x_k)$, we estimate the function g in equation (1.1) using a linear combination of the k predictors:

$$\text{score}(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k, \quad (1.2)$$

and we take the response of interest to be:

$$\Pr(\text{deception}|x) = \Pr(y|x) = \frac{e^{\text{score}(x)}}{1 + e^{\text{score}(x)}}. \quad (1.3)$$

This is technically similar to choosing $g = \text{score}(x)$, except that the random noise in equation (1.1) is now associated with the probability distribution for y in equation (1.3), which is usually taken to be Bernoulli. We are using an estimate of the score equation (1.2) as a hyperplane to separate the observations into two groups, deceptives and nondeceptives. The basic idea of separating the observations is the same for non-linear approaches. Model estimates do well if there is real separation between the two groups.

Model development and estimation for such prediction/classification models involve a number of steps:

1. Specifying the possible predictor variables (features of the data) to be used.
2. Choosing the functional form g in model (1.1) and the link function.
3. Selecting the features to be used for classification.
4. Fitting the model to data to estimate empirically the prediction equation to be used in practice.
5. Validating the fitted model through some form of cross-validation.

Different methods of fitting and specification emphasize different features of the data. Logistic regression models make no assumptions about the distribution of the predictors. The maximum likelihood methods typically used for their estimation put heavy emphasis on observations close to the boundary between the two sets of observations. Common experience with empirical logistic regression and other prediction models is that with a large number of predictor variables we can fit a model to the data (using steps 1 through 4) that completely separates the two groups of observations. However, once we implement step 5 we often learn that the achieved separation is illusory. Thus many empirical approaches build cross-validation directly into the fitting process, and set aside a separate part of the data for final testing. A thorough discussion on classification/prediction models, cross-validation, and related methodologies such as black box approaches can be found in [15].

1.4 The Data

The Department of Defense Polygraph Institute (DoDPI) provided data from 170 specific incident cases that vary by the collection agency, type of crime, formats and questions. We analyzed 149 cases*, a mix of ZCT and MGQT tests (see Table 1.2). We had to discard 21 cases due to missing information on one or more charts. The type of data missing could be any combination of type of questions, onset of the question, time of the answer, and others. All data were collected with Axciton polygraph instruments.

TABLE 1.2
Number of cases by test type and ground truth.

	Deceptive	NonDeceptive	Total
ZCT	27	24	51 (51)
MGQT	90	29	119 (98)
Total	117 (98)	53 (51)	170 (149)

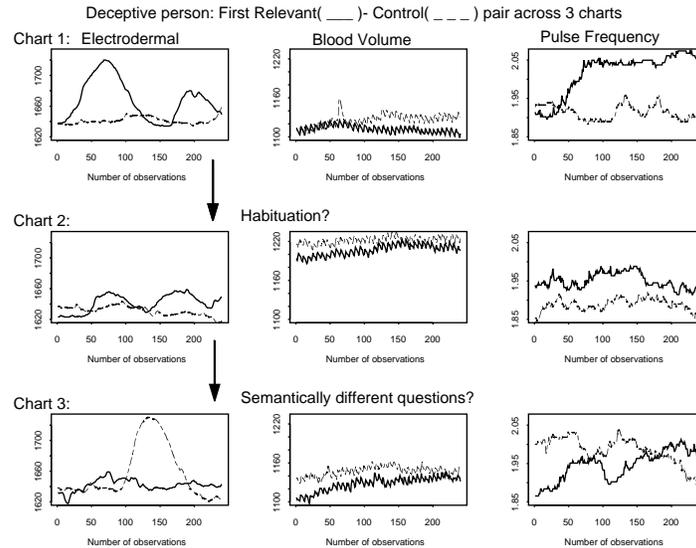
Each examination (subject/case) had three to five text data files corresponding to the exam charts, each approximately five minutes long. The data were converted to text from their native proprietary format by the “Reformat” program developed by JHUAPL. program, and the identifiers were removed. Each data file contained demographic information (when available), the sequence of questions asked during the exam, and the following fields:

1. Sample: index of observations; sampling rate is 60Hz for all measurements.
2. Time: the time, relative to the test beginning, when the sample was taken
3. Pn1: recording of the thorax respiration sensor.
4. Pn2: recording of the abdomen sensor.
5. EDR: data from an electrodermal sensor.
6. Cardio: data from the blood pressure cuff (60 and 70 mmHg inflated).
7. Event: time for the begining and end of the question and time of the answers.

Figure 1.1 shows raw data for one subject. Each time series is one of the four biological signals plus unknown error. In our analysis we use an additional series (pulse frequency that we extracted from the cardio signal)(see § 1.5.1). Demographic data such as gender, age, and education are typically available to the examiner. We had limited demographic data and did not utilize it in the current analysis.

Respiratory tracing consists of inhalation and exhalation strokes. In manual scoring the examiner looks for visual changes in the tracings for breathing rate, baseline and amplitude, where for example 0.75 inches is the desired amplitude of respiratory activity. Upper and lower respiration recordings are highly correlated as are the features we extract on these recordings.

*These data overlap with those used in the development of PolyScore

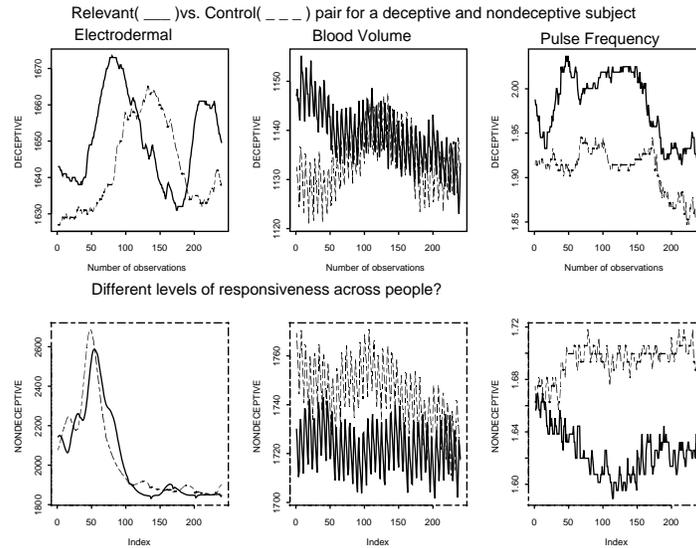
**FIGURE 1.2**

Overlaid response windows for electrodermal, blood volume and pulse frequency series on charts 1, 2 and 3 of a deceptive person for the first relevant-comparison question pair. Sampling rate here is 12Hz.

Electrodermal (EDR) tracing is the most prominent signal. Based on the limited information about the Axciton instrument, our EDR signal is a hybrid of the skin conductance and skin resistance [9] and little of known psychophysiological research can be applied. In manual scoring an evaluator may look for changes in amplitude and duration of response. When there is no reactivity the tracing is almost a horizontal line. Psychophysiology literature reports a 1-3 second delay in response to a stimulus. We observe EDR latency of 1-6 seconds from the question onset.

Research has shown that stronger stimulation elicits larger EDR response, but repetitive stimulation leads to habituation [8]. In Figure 1.2 notice the decrease in the response on the first relevant question across three charts. This could be a sign of habituation where a response to a stimulus is reduced with repeated exposure to the same question. However, in a number of cases the sequence of the questions may not be the same across the charts so we might be observing different responsiveness to semantically different questions and not habituation. For example, the first relevant question on chart 1 may appear in the third position on chart 3. In addition, different people have different responsiveness to the same stimulus (Figure 1.3).

Cardiovascular tracing records systolic stroke (pen up), diastolic stroke (pen down) and the dichotic notch. The evaluator looks for changes in baseline, amplitude, rate and changes in dichotic notch (position, disappearance). For cardiovascular activity the blood pressure usually ranges from 80mmHg to 120mmHg, but we cannot utilize this knowledge since the scale is arbitrary with respect to known physiological units.

**FIGURE 1.3**

Overlaid response windows for electrodermal, blood volume and pulse frequency series on chart 1 of a deceptive and a nondeceptive person for a relevant-comparison question pair.

In fight-or-flight situations, heart rate and blood pressure typically both increase.

Besides habituation there are other issues manifested in these data that may influence feature extraction, evaluation and modeling. There are latency differences present across different question pairs. In Figure 1.5, notice how the latency changes for the EDR as we move from the first relevant-comparison pair to the third. We can also observe different responsiveness (e.g., magnitude) across different questions. This phenomenon, however, may actually be due to a body's tendency to return to homeostasis and not due to a different reaction to different stimuli.

Our analysis revealed diverse test structures even within the same test format. The ZCT usually has the same number of comparison (C) and relevant (R) questions. A typical sequence is CRCRCR. The MGQT proposed sequence is RRCRRC. These sequences may be interspersed with other question types, and in our data we found at least 15 different sequences. The questions varied greatly across tests and were semantically different among subjects within the same crime. The order of questions varied across charts for the same person. Two problems we faced were the variable number of charts and variable number of relevant questions. Missing relevant questions should be treated as missing data; however, in this project we did not have sufficient evidence to properly impute these values. Thus we chose to drop the fourth relevant-comparison pair when it existed. For eight subjects who were missing the third relevant-comparison pair, we replaced their value by zero, i.e., we assumed that there was no difference in the response on that particular relevant-comparison pair. Elimination of both the fourth chart and the fourth relevant-comparison pair

when missing, and replacement of missing values with zeros did not significantly change the model coefficients nor the final result of classification. These types of differences across cases pose major problems for both within- and between-subject analyses, unless all the responses are averaged.

The crucial information for the development of a statistical classifier of polygraph data is ground truth (i.e., knowledge of whether a subject was truly deceptive or nondeceptive). Ideally, determination of ground truth should be independent of the observed polygraph data, although it is not clear how the ground truth was established for some of our cases. This introduces uncertainty in class labels, in particular for innocent cases since their ground truth is typically set based on someone else's confession. We proceed as though the ground truth in our data is correct.

1.5 Statistical Analysis

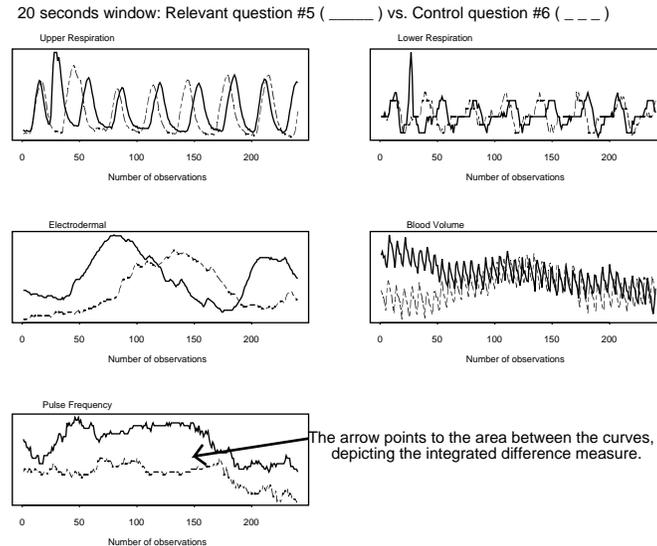
We follow the general framework described in § 1.3 for development and estimation of the logistic regression classification model. The analysis can be broken into Signal Processing, Feature Extraction, Feature Evaluation, Modeling and Classification, and Cross-Validation.

1.5.1 Signal Processing

With modern digital polygraphs and computerized systems, the analog signals are digitized and the raw digitized signals are used in the algorithm development. The primary objective of signal processing is to reduce the noise-to-information ratio. This traditionally involves editing the data (e.g., to detect artifacts and outliers), some signal transformation, and standardization. Our goal is to do a minimal amount of data editing and preserve the raw signal since we lack information on actual instrumentation and any type of filtering performed by either the machine or the examiner.

We first subsampled the 60Hz data by taking every fifth observation for each channel. Next we transformed the cardiovascular recording. We separated the relative blood volume from the pulse, constructing a new series for the relative blood pressure and another one for the pulse frequency. This was done by first calculating the average signal by applying a moving average with a window of size five. This gives a crude measurement of relative blood pressure. The averaged signal was subtracted from the original signal to produce the pulse. The pulse frequency time series is obtained by first removing ultra-high frequency by applying a low pass filter[†]. The filtered signal is made stationary by subtracting its mean. For each window of size 199 observations, we computed the spectral density of a fitted sixth order auto-regressive

[†]We used the Matlab built-in *Butterworth* filter of the 9th order at frequency 0.8.

**FIGURE 1.4**

Overlaid response windows of a relevant and a comparison question for respirations, electrodermal, blood volume, and pulse frequency series on chart 1 of a deceptive person.

model[‡]. Via linear interpolation we calculated the most prominent frequency. The procedure was repeated for the length of the time series to obtain a new series representing the frequency of a person's pulse during the exam.

1.5.2 A Simplified Approach to Feature Extraction

The discussion of general statistical methodology for prediction and classification in § 1.3 emphasized the importance of feature development and selection. A feature can be anything we measure or compute that represents the emotional signal. Our goal was to reduce the time-series to a small set of features with some relevance to modeling and classifying the internal psychological states such as deception.

Our initial analysis tried to capture low and high frequency changes of the given measurements. To capture slow changes we extracted *integrated differences* and *latency differences* features within a 20-second window from the question onset. Within the same response interval, we extracted *spectral properties differences* to capture high frequency changes. These three features are crude measures of differential activity on relevant and comparison questions. We used the same features for all signals except for the respiration where we did not use the latency.

[‡]We explored different AR models as well, but AR(6) models seem to capture the changes sufficiently.

Integrated Differences. The integrated difference is the area between two curves.

$$d_{ijkl} = \sum_{l=1}^n (R_{ijkl} - C_{ijkl}) \quad (1.4)$$

is the integrated difference of the i^{th} relevant (R) question versus the i^{th} comparison (C) question of the j^{th} channel on the k^{th} chart, where $n = 240$ is the number of observations in the response window (see Figure 1.4).

Latency Differences. We calculated latency for each 20-second window for comparison and relevant questions on all channels except respiration as follows:

1. Take the absolute value of the differenced time series, $Y_t = |\Delta X_t|$.
2. Calculate the cumulative sum, $Y_j = \sum_{k=0}^j X_k$, and normalize it, i.e., $Z_j = \frac{Y_j}{Y_n}$.
3. Define latency as the minimum Z_j such that $Z_j > 0.02$, i.e., $\ell = \min\{Z_j : Z_j \geq 0.02\}$.
4. Define the latency difference feature as the difference in the latency for the relevant and the comparison questions: $\ell_{rc} = \ell_r - \ell_c$.

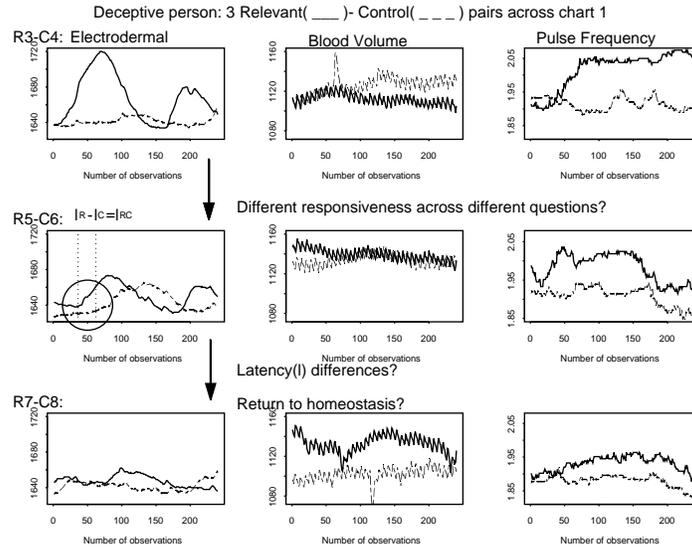


FIGURE 1.5

The overlaid response windows for 3 pairs of relevant and comparison questions on electrodermal, blood volume, and pulse frequency series on chart 1 of a deceptive person.

Spectral Properties Differences. High frequency measure is the difference between spectral properties that we defined in the following way:

1. Apply a high pass filter[§] on a 20-second window for each comparison and relevant question.
2. Generate periodograms as an estimator measure of spectrum.
3. Assess the spectral properties difference:
 - (a) Calculate a mean frequency component, $f_c = \int_0^\pi \lambda S_c(\lambda) d\lambda$, where λ is the spectral density of the process, and S_c is the estimated measure of the spectrum.
 - (b) Calculate the variance of the frequency component, $v_c = \int_0^\pi \lambda^2 S_c(\lambda) d\lambda - f_c^2$.
 - (c) Combine (a) and (b) to get $h_{rc} = |f_r - f_c| + |\sqrt{v_r} - \sqrt{v_c}|$.

These extracted features are measures of responsiveness to the stimuli. For integrated and latency differences measures we expect positive values if the response is higher on the relevant question, negative if it's higher on the comparison questions and zero if there is no difference. Spectral properties differences only give the magnitude of the differential activity.

1.5.3 Feature evaluation, modeling, and classification

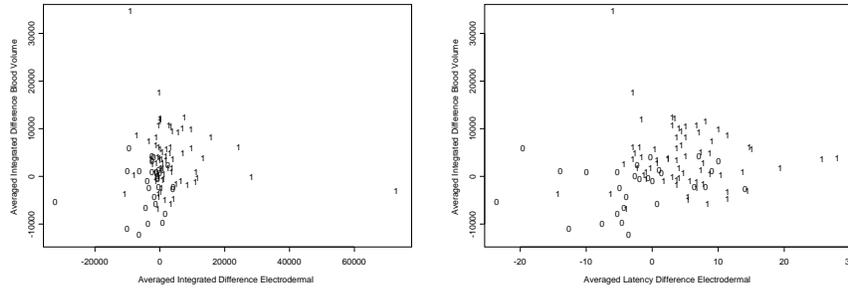
This section reviews aspects of feature selection and of statistical modeling involving the development of scoring rules into classification rules. The extracted features were considered in three types of comparisons between relevant and comparison questions:

1. each relevant compared to its nearest comparison,
2. each relevant compared to the average comparison,
3. averaged relevant compared to the average comparison.

In the first two settings the maximum number of continuous variables per subject was 240 (4 relevant-comparison pairs \times 5 channels \times 4 charts \times 3 features), while the third setting had 60. Since the potential variable space is large relative to the sample size and the variables are highly correlated, particularly in the first two settings, we evaluated features graphically, via clustering, principal-component analysis (PCA) and with univariate logistic regression trying to reduce dimensionality. The remainder of this report will focus on setting 3. The other two settings are briefly discussed in [24].

Figure 1.6 shows separation of the two classes given the integrated difference feature for the electrodermal (dEdr) channel or the electrodermal latency difference (lEdr) versus the integrated difference for blood volume (dBv). These are values averaged across charts. Most deceptive subjects (the 1s in the figure) have values greater than zero on dEdr and their distribution is slightly skewed left on pulse frequency (Fqp). Nondeceptive subjects (represented with 0s) mostly have values less than zero on dEdr. They are less variable on Fqp and are centered around zero. Most deceptive subjects have a positive latency difference measure; their latency is

[§]We used built-in *Butterworth* filter from Matlab.

**FIGURE 1.6**

Bivariate plots of averaged integrated difference feature for electrodermal (dEdr) versus averaged integrated difference feature for blood volume (dBv), averaged latency difference for electrodermal(IEdr) versus dBv

longer on relevant than on comparison questions (when averaged within and across charts). Nondeceptives show a tendency of having less variable values that are less than zero (i.e., longer latency on EDR on comparison questions, but not as much between variability as for deceptive subjects). A bivariate plot of the integrated difference for blood volume and pulse frequency shows less clear separation of the two groups. Deceptive subjects show a tendency to have higher blood volume responses on relevant than on comparison questions while the opposite holds for nondeceptive examinees.

1.5.4 Logistic Regression

We used data from 97 randomly chosen subjects (69 deceptive, 28 nondeceptive) for evaluation and selection of the best subset of features for the classification model, and saved the remaining 52 cases (29 deceptive, 23 nondeceptive) for testing. Since the questions across subjects are semantically different and there is no consistent ordering we developed a model based on comparison of the averaged relevant questions versus the averaged comparison questions. Logistic regression models developed on variables from the first two settings even when principal components are used as predictors yield multiple models with at least 9 predictors. These predictors vary across different models and perform poorly on the test set, although they may achieve perfect separation on the training dataset [24].

Average Relevant vs. Average Comparison. For each chart, each channel and each feature we calculated the average relevant and average comparison response over the 20-second window. Typically, if we have a noisy signal, one simple solution is to average across trials (i.e. across charts) even though we lose some information on measurement variability between different charts.

$\bar{R}_{ij.} = \frac{\sum_{k=1}^{nr_i} R_{ijk}}{nr_i}$ is the averaged relevant response and $\bar{C}_{ij.} = \frac{\sum_{k=1}^{nc_i} C_{ijk}}{nc_i}$ is the averaged comparison response on the i^{th} chart, j^{th} channel, where nr_i is the number of

relevant questions and nc_i is the number of comparison questions on the i^{th} chart. We calculate the averaged relevant ($\bar{R}_{.j}$) and comparison ($\bar{C}_{.j}$) responses across m charts producing a total of 13 predictors: 5 for integrated differences, 5 for spectral proportion differences and 3 for latency differences.

The logistic regression was performed for each feature independently on each chart, across the charts and in combination to evaluate the statistical significance of the features. A stepwise procedure in Splus software was used to find the optimal set of features. Neither clustering nor PCA improved the results. The following models are representative of performed analyses on each feature and when combined: Integrated Difference (M1), Latency Difference (M2), Spectral Properties Difference (M3): $\text{Score} = \hat{\beta}_0 + \hat{\beta}_1 \text{hPn1} + \hat{\beta}_2 \text{hEdr} + \hat{\beta}_3 \text{hBv} + \hat{\beta}_4 \text{hFqp}$, and All 3 features (M4).

TABLE 1.3

Features with the estimated logistic regression coefficients and standard errors for models M1, M2 and M4.

Model	M1	M2	M4
Features	$\hat{\beta}$ (SE)	$\hat{\beta}$ (SE)	$\hat{\beta}$ (SE)
Intercept $\times 10$	+4.90 (3.03)	+6.80 (2.50)	-3.07 (2.96)
Integrated Diff. Electrodermal $\times 10^4$	+3.15 (1.02)		+1.59 (0.62)
Integrated Diff. Blood Volume $\times 10^4$	+2.14 (0.75)		+1.07 (0.44)
Integrated Diff. Pulse Frequency $\times 10$	-3.72 (1.44)		-2.49 (0.87)
Latency Diff. Electrodermal $\times 10$		+1.43 (0.404)	+0.35 (0.38)
Spectral Diff. Blood Volume $\times 10$			+0.36 (0.23)

We considered models on individual charts and observed almost identical models across charts. Chart 3 did worse on cross-validation than the other two, and relied more on high frequency measures of respiration and sweating. Chart 2 added significantly to the detection of innocent subjects in comparison to chart 1. For chart 2 the latency difference on blood volume was a slightly better predictor than the high frequency measure which is more significant on chart 1 [24].

The linear combination of integrated differences was the strongest discriminator. Latency had the most power on the electrodermal response. Our high frequency feature on any of the measurements was a poor individual predictor, particularly on nondeceptive people, however it seems to have some effect when combined with the other two features. All features show better discrimination on electrodermal response, blood volume and pulse frequency than on the respiration measurements.

1.5.5 Classification Results

We tested the previously described models for their predictive power on an independent test set of 52 subjects. Table 1.4 summarizes the classification results based on a 0.5 probability cutoff. A probability of 0.5 or above indicates deception, and a probability less than 0.5 indicates truthfulness.

We ran the same subsets of training and test data through the Polyscore. Figure 1.7

TABLE 1.4

Percents of correctly classified subjects in hold-out-set cross-validation.

Model	Training		Test	
	Deceptive(%)	Nondeceptive(%)	Deceptive(%)	Nondeceptive(%)
M1	94	64	97	52
M2	96	29	90	9
M3	99	7	100	9
M4	93	61	97	48

shows receiver operating characteristic curves (ROCs) of model M4 performance and of PolyScore 5.1 on 52 test cases. This is not an independent evaluation of PolyScore algorithm since some of these test cases were used in its development. ROC and the area under the curve give a quantitative assessment of a classifier's degree of accuracy. [7] showed that ROC overestimates the performance of the logistic regression classifier when the same data are used to fit the score and to calculate the ROC.

TABLE 1.5

5-fold cross-validation results at 0.5 probability cutoff value.

	Training (N=119)				Test(N=30)			
	Deceptive%		Nondeceptive%		Deceptive%		Nondeceptive%	
	M1	M4	M1	M4	M1	M4	M1	M4
Mean	91	90	60	57	92	92	70	66
St.Error	4.1	2.1	2.3	2.1	5.4	5.6	21.6	21.5

Since k -fold cross-validation works better for small data sets [15] we performed 5-fold cross-validation on models M1 and M4. The results are presented in Table 1.5. The number of innocent subjects in training runs vary from 39 to 47 out of 51, and deceptive from 72 to 80 out of 98. In the first run with only four nondeceptive test cases our models achieve 100% correct classification which is highly inconsistent with the other runs. The average area under the ROC for M4 is 0.899(± 0.05). When we apply shrinkage correction proposed by [7] the average area under the curve is approximately 0.851.

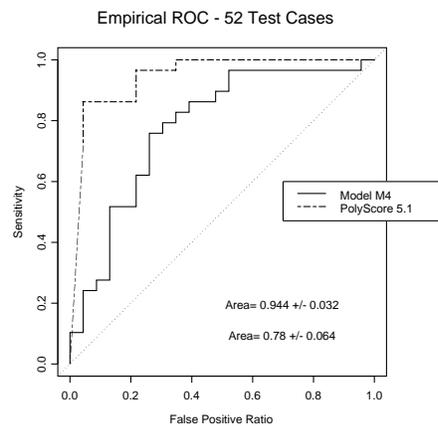
1.6 Discussion

The objective of the automated scoring algorithms for polygraph data is to create reliable and statistically valid classification schemes minimizing both false positive and false negative rates. Beginning in the 1970s, various papers in the polygraph literature have offered evidence claiming to show that automated classification methods for analyzing polygraph charts could do so. According to [9], the accuracies of five different computer algorithms range from 73% to 89% on deceptive subjects when inconclusives are included and 91% to 98% when they are excluded. For innocent

subjects these numbers vary from 53% to 68%, and 72% to 90%.

Our analyses based on a set of 149 criminal cases provided by DoDPI suggest that it is easy to develop such algorithms with comparable recognition rates. Our results are based on a 0.5 probability cutoff for two groups: deceptive (probability greater than 0.5) and nondeceptive (probability less than 0.5). Other cutoff values would allow us to balance the errors differently. Neither clustering nor PCA significantly improve our results, which is consistent with the recent work of [5].

One possible explanation for the relatively poor classification performance is the small sample size and, in particular, the small number of nondeceptive cases. However, PolyScore algorithms, for example, had a much larger database [22, 12], but their accuracy rates are not significantly better than ours.



≥ 0.5 cutoff	M4	PolyScore 5.1	Others*
Deceptive	90%	86%	73-100%
NonDeceptive	48%	78%	53-90%

FIGURE 1.7

ROCs for classification results of PolyScore 5.1 and M4 on 52 test cases. The table shows percent correct when 0.5 is a cutoff value. In practice PolyScore 5.1 uses 0.95 and 0.05 as the cutoff values for classifying deceptive and nondeceptive subjects. (*)These values are based on different cutoffs and range over the results when inconclusives are included and excluded giving higher percent correct when inconclusive cases are excluded.

Another possible explanation could be high variability and presence of measurement errors that come with real-life polygraph data, where there is a lack of standards in data collection and recording. Our exploratory data analysis points to problems with question inconsistency, response variability within an individual and across individuals (due to nature, gender, etc.), and possible learning effects. It is not always clear where differences in responses comes from; are we dealing with habituation

or comparing semantically different questions across the charts and hence having different responsiveness to different questions? Since in our data the questions are semantically different, and no consistent ordering within and across charts could be established, we averaged the relevant and comparison responses and then look at their difference. PolyScore and CPS algorithms take the same approach. This methodology ignores the question semantics which could be a flaw in our approach. These phenomena could be better studied in the screening type tests or with more standardized laboratory cases[¶] where there is consistency in the order and the type of questions asked within and across both charts and individuals. Although CPS algorithms have been developed on laboratory data, they have not achieved significantly better results.

These points and numerous aspects of each of the steps of the analysis are discussed further in [24]. Perhaps it is not reasonable to expect that a single algorithm will successfully be able to detect deceptive and truthful examinees. A solution may lay in the data collection and on detailed research on underlying theory for polygraphs, before proper statistical modeling can be effectively utilized.

Finally, we note that in the cases we examined there is little or no information available on the assessment of ground truth, differences among examiners, examiner-examinee interactions, and delays in the timing of questions. Most of these are not addressed by current scoring algorithms. More discussion on these issues and their implications for inflated accuracies can be found in Appendix G of [6].

1.7 Conclusion

This paper presents an initial evaluation and analysis of polygraph data for a set of real-life specific incident cases. With a very simple approach, in a short period of time, we have managed to obtain accuracy rates to a certain degree comparable to what's currently being reported by other algorithms and manual scoring. The fact that we are able to produce a number of different models that account for different predictors yet give similar results, points to the complexity that underlines assessment and/or classification of examinee's deceptiveness.

This work can be redefined and extended in a number of ways. More features could be extracted and explored. Thus far these efforts have not resulted in significantly smaller errors, hence it raises a question how far could this approach go? One could imagine improvements to current methods by running a proper Bayesian analysis and incorporating prior knowledge on prevalence. Our inclination would be to do a more complex time series analysis of these data. The waveform of each channel can be considered and the analysis would gear towards describing a physiological signature for deceptive and nondeceptive classes. Clearly the ordering of

[¶]See discussion on possible downfalls of lab data in [6].

the questions should be accounted for. A mixed-effects model with repeated measures would be another approach, where repetitions would be measurements across different charts. In other areas with similar data researchers have explored the use of Hidden Markov Models [10].

There has yet to be a proper independent evaluation of computer scoring algorithms on a suitably selected set of cases, for either specific incidents or security screening, which would allow one to accurately assess the validity and accuracy of these algorithms. One could argue that computerized algorithms should be able to analyze the data better because they execute tasks which are difficult even for a trained examiner to perform, including filtering, transformation, calculating signal derivatives, manipulating signals, and looking at the bigger pictures, not merely adjacent comparisons. Moreover, computer systems never get careless or tired. However, success of both numerical and computerized systems still depends heavily on the pre-test phase of the examination. How well examiners formulate the questions inevitably affects the quality of information recorded. We believe that substantial improvements to current numerical scoring may be possible, but the ultimate potential of computerized scoring systems depends on the quality of the data available for system development and application, and the uniformity of the examination formats with which the systems are designed to deal. Computerized systems have the potential to reduce the variability that comes from bias and inexperience of the examiners and chart interpreters, but the evidence that they have achieved this potential is meager at best.

Acknowledgments

The author would like to thank Stephen E. Fienberg and Anthony Brockwell, for their advice and support on this project. The author is also grateful to the members of the NAS/NRC Committee to review the scientific evidence on polygraph for the opportunity to work with them, and to Andy Ryan and Andrew Dollins of Department of Defense Polygraph Institute for providing the data.

References

- [1] Alder, K. 1998. To Tell the Truth: The Polygraph Exam and the Marketing of American Expertise. *Historical Reflections*. 24(3), pp.487-525.
- [2] American Polygraph Association. www.polygraph.org
- [3] Brownley, K.A., B.E. Hurwitz, N. Schneiderman. 2000. Cardiovascular psychophysiology. Ch. 9, pp. 224-264, in [4].
- [4] Cacioppo, J.T., Tassinari, J.T., Bernston, G.G., Eds. 2000. *Handbook of Psychophysiology*. Second Edition. New York: Cambridge University Press.

- [5] Campbell, J.L. 2001. *Individual Differences in Patterns of Physiological Activation and Their Effects on Computer Diagnoses of Truth and Deception*. Doctoral Dissertation. The University of Utah.
- [6] Committee to Review the Scientific Evidence on the Polygraph. 2002. *The Polygraph and Lie Detection*. Washington, DC: National Academy Press.
- [7] Copas, J.B., Corbett, P. 2002. Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika*, 89(2), pp. 315-331.
- [8] Dawson, M., A.M. Schell, D.L. Filion. 2000. The electrodermal system. Ch. 8, pp. 200-223, in [4].
- [9] Dollins, A.B., D.J. Kraphol, D.W. Dutton. 2000. Computer Algorithm Comparison. *Polygraph*, 29(3), pp.237-257.
- [10] Fernandez, R. 1997. *Stochastic Modeling of Physiological Signals with Hidden Markov Models: A Step Toward Frustration Detection in Human-Computer Interfaces*. Master's Thesis. Massachusetts Institute of Technology.
- [11] Gratton, G. 2000. Biosignal Processing. Ch. 33, pp. 900-923, in [4].
- [12] Harris, J.C., Olsen, D.E. 1994. Polygraph Automated Scoring System. U.S. Patent #5,327,899.
- [13] Harris, J. 1996. Real Crime Validation of the PolyScore 3.0 Zone Comparison Scoring Algorithm. Johns Hopkins University Applied Physics Laboratory.
- [14] Harver, A., T.S. Lorig. 2000. Respiration. Ch. 10, pp. 265-293, in [4].
- [15] Hastie, T., R. Tibshirani, J. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag.
- [16] Hosmer, D.W., Lemeshow, Jr. S. 1989. *Applied Logistic Regression*. New York: John Wiley & Sons.
- [17] Jennings, R.J., L.A. 2000. Salient method, design, and analysis concerns. Ch. 32, pp. 870-899, in [4].
- [18] Kircher, J.C., and D.C. Raskin. 1988. Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology* 73:291-302.
- [19] Kircher, J.C., and D.C. Raskin. 2002. Computer methods for the psychophysiological detection of deception. Chapter 11, pp. 287-326, in *Handbook of Polygraph Testing*, M. Kleiner, ed. London: Academic Press.
- [20] Matte, J.A. 1996. *Forensic Psychophysiology Using Polygraph-Scientific Truth Verification Lie Detection*. Williamsville, NY: J.A.M. Publications.
- [21] McLachlan, G.J. 1992. *Discriminant Analysis and Statistical Pattern Recognition*. New York: John Wiley & Sons.
- [22] Olsen, D.E, Harris, J.C., Capps, M.H., Ansley, N. 1997. Computerized Polygraph Scoring System. *Journal Forensic Science* 42(1):61-71.
- [23] Raskin, D.C., Hont, C.R., Kircher, J.C. 1997. The scientific status of research on polygraph techniques: The case for polygraph tests. In D.L. Faigman, D.Kaye, M.J. Saks, J. Senders (Eds.) *Modern scientific evidence: The law and science of expert evidence*. St.Paul, MN:West.

- [24] Slavkovic, A. 2002. *Evaluating Polygraph Data*. Technical report 766. Department of Statistics. Carnegie Mellon University.
- [25] Swinford, J. 1999. Manually Scoring Polygraph Charts Utilizing the Seven-Position Numerical Analysis Scale at the Department Of Defense Polygraph Institute. *Polygraph*, 28(1), pp.10-27.