ORNL-13114

Computer Science and Mathematics Division

Mathematical Sciences Section

LARGE DATASETS: SEGMENTATION, FEATURE EXTRACTION, AND COMPRESSION

D. J. Downing
V. Fedorov
W. F. Lawkins
M. D. Morris
G. Ostrouchov

Mathematical Sciences Section
Computer Science and Mathematics Division
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, Tennessee 37831-6367

Date Published: July 1996

Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed for the United States Department of Energy by Martin Marietta Energy Systems, Inc.

Prepared by the
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831
managed by
Lockheed Martin Energy Research Corp.
for the
U.S. DEPARTMENT OF ENERGY
under Contract No. DE-AC05-96OR22464

Contents

List of Figures

List of Tables

LARGE DATASETS:
SEGMENTATION, FEATURE EXTRACTION, AND COMPRESSION

D. J. Downing

V. Fedorov

W. F. Lawkins

M. D. Morris

G. Ostrouchov

Abstract

Large data sets with more than several million multivariate observations (tens of megabytes or gigabytes of stored information) are difficult or impossible to analyze with traditional software. The amount of output which must be scanned quickly dilutes the ability of the investigator to confidently identify all the meaningful patterns and trends which may be present. The purpose of this project is to develop both a theoretical foundation and a collection of tools for automated feature extraction that can be easily customized to specific applications. Cluster analysis techniques are applied as a final step in the feature extraction process, which helps make data surveying simple and effective.

1. Introduction

The primary problem considered in this study is that of identifying “unusual” segments of data from very long streams of scalar, or univariate data series, where the adjective “unusual” is intended to convey a sense that the fraction of unusual to usual segments over the whole data stream is small. In this study we also consider the problem of partitioning unusual segments into similarity classes. The assumption is that the collection of all unusual segments, or at least each of the similarity classes, is amenable to traditional and more detailed types of statistical analysis and that it is practical to do so because collectively they are a small fraction of the whole data stream. These problems are viewed here both from theoretical and pragmatic standpoints. Sample analyses using model data are presented to help explain the theoretical ideas, and sample analyses of data series measurements of two very different processes using four variants of the general class of methods described in this study are also presented.

Note the use of the expression “data series” as opposed to “time series” in the above discussion. That usage is intended to emphasize the idea that the data frequently does not admit a natural sense of “future” and “past.” Even if the physics of the process being observed does allow a sense of time to be identified with the data series, the analysis techniques we consider are different from traditional time series methods that focus on “forecasting ” or “online” analyses. In general, the methods we consider compare a segment to a collection of its neighbors. If there is a sense of time associated with the data series, that collection includes segments observed both earlier and later than the one being analyzed. In order to give meaning to “neighbor”, some sense of data ordering is essential for the methods we discuss in this study.

The approach explored here for attacking the two problems described above is to first partition the data series into relatively short segments and then model each segment using a relatively simple, low-order model of some type. If the data series segmentation is a wise choice, then ordinary segments will be well represented by the simple model and, conversely, unusual segments will be identified as ones that have to be represented by a more sophisticated model. In general, the parameters of the model are expected to have typical values and not display significant variation over the collection of usual segments. However, over the collection of unusual segments the model parameters are expected to vary significantly. If the unusual segments are reflections of a limited number of distinct digressions from the process corresponding to the usual segments, then the relatively large variation in model parameter values over the whole collection of unusual segments may cluster into a few classes of relatively small variation. From a statistical standpoint, we talk about a mixture of populations of different sizes and the detection and segregation of those populations.

The two problems being addressed in this study can be viewed as pieces of the more general problem of analyzing large data sets. The methods investigated here can, in principle, be generalized to those large data sets that admit a meaningful segmentation. The most obvious example is a very long, high-dimensional time series. In order to use the ideas and techniques discussed here for that larger problem, two steps have to be taken. First, the high-dimensional data series has to be projected to a low-dimensional data series, and second, the methodology described in this study has to be extended to include low-dimensional, cross-correlated multivariate data series. Actually, the latter step is conceptually straightforward for our approach. The difficulties are technical in nature and are mostly confined to developing fast algorithms for implementing methods.

Let us consider what should be the most important aspects of the problems we have determined to study. Thinking in terms of some process, our purpose may be to identify unusual temporal events that are part of the process, or it may be to identify perturbations relative to that process. In either case, the primary objective is to locate something unusual rather than to locate and accurately classify an unusual event. Discussions of accuracy in the context of the problems we are considering mean only that, first, we can say something about the percentage of perturbations that are missed, and that, second, we can say something about measures used to divide perturbations into broad, but distinct, categories. Accuracy in the usual statistical sense of either inference or prediction is not as important as is speed of calculation. This is why we prefer simple, low-dimensional models.

Consider how to segment the data. This aspect of the location problem is both critical and application dependent. We assume the scientist, engineer, biologist, or whoever is using these methods has an idea about the nature, including space and time scales, of the events or perturbations of importance for their particular application. We have two distinct situations if one application is focused on a single time scale, whereas another application includes a wide range of time scales. In the former case, we may choose a segmentation size such that a corresponding partition of the data is expected, for a large data set, to separate almost all events into distinct segments. Still, even in this situation we want the usual segments to be amenable to a simple model such that the sense of "usual" can be quantified by means of a relatively small variation in model parameters. In the latter case, the same segmentation strategy would force us to choose a size sufficiently large to accommodate the largest time scale of interest. However, a partition based on that size would not be expected to segregate the large majority of unusual events nor would it be likely to admit a simple model for usual events. If a single value for size is to be used, then size is likely, in fact, to be dictated by the upper bound of what can be classified as both usual and simple. Evidently there are applications for which a multistage, or hierarchical segmentation structure is most appropriate, especially in the case of self-similar processes (for definitions and discussion, see [5, 20, 28]). The techniques discussed here can be applied at every segmentation level. Developing techniques for defining hierarchical segmentation structures is a very interesting topic for future studies.

Most of the theoretical material included in this article is described in terms of regression models. However, four distinct methods for detecting unusual segments, or change points, in data series are presented, and the performance of each technique is examined by applying it to two example time series, one representing an atmospheric process and the other a neurophysiological process. The atmospheric process is represented by a time series collected under the auspices of the Atmospheric Radiation Measurement (ARM) project [11]. The time series used here is measurement of the liquid water content of the atmosphere at a U.S. location. The background process in this case corresponds to a relatively clear day with dry conditions. Perturbations include cloud, rain, and fog events as well as some instrument malfunction events. The second time series is a segment taken from one channel of a sixteen channel electroencephalogram (EEG) record for an epileptic patient [18]. EEG records typically include a great deal of "artifact," representing head movement, eye movement, muscle tension, grinding teeth, etc., in addition to unmasked neurophysiological activity. If we associate neurophysiological activity with the background process, than artifact is a perturbation relative to that background process.

The remainder of this article is organized as follows. In Section 2 we briefly examine the problem of data segmentation. In Section 3 we review local regression models and examine those theoretical properties that can be exploited to detect unusual events. This is done largely by viewing each member of the selected segmentation independently. In Section 4 several theoretical means for relaxing the condition that each member of the segmentation is independent of all others is examined. Section 5 addresses the issue of optimal basis function selection. It also relates regression modeling and empirical basis construction to the data analysis approach inspired by the theory of nonlinear dynamical systems. Section 6 presents four perturbation, or change point, detection techniques together with applications of those techniques to two example time series. Section 7 includes a summary of our findings.

2. Segmentation of the Data

We start with the univariate case. The generalization to the multivariate case demands more sophisticated techniques, but nevertheless is based on the same ideas.

Let $\{y_i\}_{i=1}^N$ be observations made under conditions $\{x_i\}_{i=1}^N$. We do not make any special assumptions about the conditions at this stage. Sometimes we consider the simple conditions $\{i\}_{i=1}^N$ rather than the more general $\{x_i\}_{i=1}^N$.

The essence of the problem is that N is very large and, therefore, it may be technically difficult to manipulate the set $Y = \{y_i\}_{i=1}^N$ or even to store it on a relatively small computing platform. In this study we consider methods based on partitioning Y into J subsets of equal size, so that

$$Y = \cup_{j=1}^J Y_j, \quad Y_j = \{y_i\}_{i=(j-1)L+1}^{jL}, \tag{1}$$

where

$$J \ll N, \quad N = L \times J,$$

and L is the size of the subset Y_j. We have to find a few descriptors for Y_j and store or work with them. Evidently, some "physical" ideas must support those descriptors, like, for instance, the parameters of a short-term model of the process that generated Y. In general, the partition size L must also depend on some physical properties of the process that generated Y. When the subscript i stands for the time y_i was generated, or observed, then a natural choice of the partition size L may be, for example, day, week, month, or half-time of decay.

A natural generalization of the segmentation scheme Eq. (1) is to use a sliding window of length L, so that

$$Y = \cup_{j=1}^J Y_j, \quad Y_j = \{y_i\}_{i=j}^{j+L-1}, \tag{2}$$

where

$$J = N - L + 1.$$

From a theoretical standpoint, the major difference between the two segmentation schemes is that the former Eq. (1) is a partition and the latter Eq. (2) is not a partition. From a pragmatic standpoint based on computational requirements, the major difference is that for the segmentation Eq. (1) we have $J \ll N$ while for the segmentation Eq. (2) we have $J \approx N$.

We realize the discussion to this point provides no guidance for choosing the segments Y_j and that this is an important topic. Clearly, the choice depends not just on the data, but also on the information we want to glean from that data. An objective function to use in an optimization scheme for this purpose could depend on a variety of different parameter types. For example, parameters to prescribe how accurately the processes we want to preserve are modeled, other parameters concerning the amount of data we are willing to store, and how fast the data has to be processed. Such an optimization technique could be a valuable contribution to information theory or large data set processing theory. However, at this stage of the development of theory and techniques for analyzing large data sets, we are mainly concerned with the mathematical tools used to process the data given a partition.

3. Local Regression Models

Regression analysis is one of the most developed areas in statistics, both from the theoretical and computational standpoints. The principal idea is to separate the description of the observed variable into the sum of a deterministic component and a stochastic component. The deterministic term is used to describe functional dependencies that we know and understand relatively well. The stochastic component is a simple representation for either what we only vaguely understand or what would require a very detailed deterministic model. The regression models considered here are linear and, thus, depend on computational linear algebra methods for implementation. These linear algebra methods are well understood and have very efficient software implementations (see LAPACK [2]). In spite of their simplicity, regression models are well suited to detecting interesting features in segmented data.

3.1. Definitions and Notation

Let

$$x_{j0} \in X_j = \{x_i\}_{i=(j-1)L+1}^{jL},$$

and suppose qualitatively speaking that x_{j_0} is not close to one of the boundary points of the interval X_j. Also, define

$$u_{ji} = x_i - x_{j_0}.$$

We are going to study the local regression model

$$y_i = \theta_j^T f(u_{ji}) + \gamma_j^T \phi(u_{ji}) + \epsilon_i, \tag{3}$$

where $(j-1)L+1 \leq i \leq jL$. In this regression model, the first term $\theta_j^T f(u_{ji})$ describes the "standard" component of the data set Y_j, and the second term $\gamma_j^T \phi(u_{ji})$ describes occasional perturbations, like contaminations, outliers, thresholds, etc. Further, the expressions $f(u)$ and $\phi(u)$ are vectors of known functions, $\theta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^q$ are unknown parameters, and ϵ_i represents whatever can not be modeled better than by a "random noise" model; i.e., ϵ_i are random variables. Concerning the random noise ϵ_i, we assume

$$E(\epsilon_i) = 0, \quad E(\epsilon_{i_i}, \epsilon_{i_2}) = \sigma_j^2 \delta_{i_1 i_2},$$

where $\delta_{i_1 i_2} = 1$, for $i_1 = i_2$ and is zero otherwise. The parameters θ_j, γ_j, σ_j^2, or some functions of them, form the set of descriptors to be saved. Generally, $p \ll L$ and $q \ll L$, meaning that the vectors θ_j, γ_j are small relative to the data Y_j. In some cases, the practitioner is interested only in perturbed intervals X_j, in which case only γ_j may be saved. As perturbed intervals are, in general, expected to be the exception, this can result in another significant reduction in saved information.

3.2. Regression Analysis Based on the Standard Term

Usually, the term $\theta_j^T f(u_{ji})$ is a low-order piece of some expansion, say a Taylor or Fourier expansion, or an expansion with respect to the eigenfunctions of some operator that, at least approximately, describes the data. A practitioner generally likes to work mainly with this term and considers the perturbation term $\gamma_j^T \phi(u_{ji})$ as a nuisance component. Besides, we expect $\gamma_j^T \phi(u_{ji}) \neq 0$ to be the exception rather than the rule. Therefore, we can work with the simpler regression model regularly and test, for instance, the hypothesis $H_0 : \gamma_j = 0$. If that hypothesis is rejected, than we can proceed with the more complicated model or just mark the corresponding intervals and save them for subsequent analysis.

The following assumptions allow us to use some well developed statistical tools.

Assumption 1. *The random variables ϵ_i are independent and normally distributed, with $E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = \sigma^2$.*

Assumption 2. *For two distinct values of j, there is no relationship between the two pairs of parameter values (θ_j, γ_j).*

These assumptions are rather strong. Concerning Assumption 1, it is generally good statistical practice to begin with normality assumptions to establish a theoretical and algorithmic foundation. Particularly for large data sets, we expect the results can be generalized to other distributions due to the Central Limit Theorem or that they may be reformulated in terms of first and second moments, and in either case that this can be done in a straightforward manner. Concerning Assumption 2, the most significant interactions are likely to occur for neighboring segments as a consequence of boundary effects. We assume the segments are sufficiently large to neglect those boundary effects, even in multidimensional cases. By neglecting the possibility of any type of coupling between distinct pairs of intervals X_j, we will make less certain inferences. For example, the probability of rejecting the null hypothesis $H_0 :\ \gamma_j = 0$ when it is wrong may be greater than the same probability under more realistic assumptions. Again, we begin with these assumptions and explore where they lead, after which we can consider the consequences of various modifications.

The model described by Eq. (3) with Assumptions 1 and 2 is a typical regression model. In the framework of this model, the parameters θ_j are the descriptors of the j-th interval X_j. To account for the presence of perturbations, we can use the descriptor pair (θ_j, γ_j). When $\gamma_j = 0$, the best linear unbiased estimator for θ_j under Assumption 1 is also the maximum likelihood estimator

$$\hat{\theta}_j = M_{jff}^{-1} \sum_i f(u_{ji}) y_i, \tag{4}$$

where

$$M_{jff} = \sum_i f(u_{ji}) f^T(u_{ji}) \tag{5}$$

and $1 \le i \le L$ (see [24, 27]). When it does not lead to a misunderstanding, we will omit the subscript j. We note that for the model Eq. (3) the maximum likelihood estimator is equivalent to the least squares estimator.

From Eq. (4) it follows that

$$E(\hat{\theta}_j) = \theta_j + M_{jff}^{-1} M_{jf\phi} \gamma_j, \tag{6}$$

where

$$M_{jf\phi} = \sum_i f(u_i)\phi(u_i).$$

If all intervals X_j have the same length and the point x_{j0} is in the same position within X_j, for all j, then the matrices M_{jff} and $M_{jf\phi}$ are the same for all j. If that is the case, we can suppress the subscript j and use M_{ff} and $M_{f\phi}$ in all subsequent formulae, and what is more important, this results in a very significant reduction in computing requirements. The reader may proceed with generalizations of these results for intervals X_j of varying length in a straightforward manner.

From Eqs. (4),(6) we find that

$$\begin{aligned} D = \mathrm{Var}(\hat{\theta}_j) &= E[(\hat{\theta}_j - E(\hat{\theta}_j))(\hat{\theta}_j - E(\hat{\theta}_j))^T] \\ &= \sigma^2 M_{ff}^{-1} \end{aligned} \tag{7}$$

and

$$\begin{aligned} R &= E[(\hat{\theta}_j - \theta_j)(\hat{\theta}_j - \theta_j)^T] \\ &= \sigma^2 M_{ff}^{-1} + M_{ff}^{-1} M_{f\phi} \gamma_j \gamma_j^T M_{\phi f} M_{ff}^{-1}. \end{aligned} \tag{8}$$

Note that σ^2 is assumed to be known in Eqs. (7) and (8).

The estimator $\hat{\theta}_j$ is insensitive to the presence of nonzero γ_j only if (see Eqs. (6) and (8)) the vector functions $f(u)$ and $\phi(u)$ are orthogonal; i.e., $M_{f\phi} = 0$. Actually, at this stage we have a rather interesting situation. In most statistical problems it is necessary to minimize the bias $M_{ff}^{-1} M_{f\phi} \gamma_j$. This objective generates a number of approaches associated with the "robust estimator" paradigm. In the present case the detection of a perturbation may be of primary importance. If that is the case and the value of $\hat{\theta}_j$ is used to identify $\gamma_j \neq 0$, then the orthogonality condition on the functions f and ϕ is not very useful. Intuitively, variation in $\hat{\theta}_j$ appears to be a good indicator that $\gamma_j \neq 0$, however there exist better indicators that are discussed in the next section.

3.3. Detection of a Perturbation using Information from a Single Interval

If the variance of ϵ_i is known, then statistical theory (see [24, 27]) tells us

$$\nu^2(\hat{\theta}_j) = \min_{\theta} \sum_i \sigma^{-2}(y_i - \theta f(u_{ji}))^2 \tag{9}$$

should be used as a statistic to detect the presence of perturbations. If the observed value of $\nu^2(\hat{\theta}_j)$ is greater than a prescribed level $\chi^2_{1-\alpha}$, then the hypothesis H_0: $\gamma_j = 0$ is rejected. In our notation, α is the probability of falsely rejecting the null hypothesis H_0. In classical statistics, choosing an appropriate value of α is a largely subjective decision. However, in the present case α has a simple and meaningful interpretation, namely, the fraction of usual segments that are falsely analyzed to be unusual. A sense of cost associated with a particular choice of α can be easily assigned. In the event $\gamma_j = 0$, the random variable $\nu^2(\hat{\theta}_j)$ has the central χ^2 distribution with $L - p$ degrees of freedom, whereas if $\gamma_j \neq 0$ it has the non-central χ^2 distribution with the same number of degrees of freedom and with the non-centrality parameter value

$$\Delta = \sigma^{-2}\gamma_j^T(M_{\phi\phi} - M_{f\phi}M_{ff}^{-1}M_{f\phi})\gamma_j. \tag{10}$$

Given L, γ_j, f, and ϕ, the parameter Δ defines the power $1-\beta$ of the χ^2 test, where β is the probability of falsely rejecting the hypothesis H_1: $\gamma_j > 0$. We note that β is a monotonically decreasing function of Δ.

We can not control the values γ_j, f, and ϕ, but the interval length L may be optimized as

$$L^* = \arg\max_L \frac{\Delta(L)}{\sqrt{L-p}}, \tag{11}$$

where "arg" stands for an arbitrary one of the possible solutions of the optimization problem. The form of Eq. (11) may be understood if we take into account the relationships

$$E[\Delta(L)] \quad = \quad L - p + \Delta(L),$$

$$\mathrm{Var}[\Delta(L)] \;=\; 2(L-p). \tag{12}$$

Usually, $L \gg p$.

The following example is an analytic demonstration for the estimation of an optimal segmentation size Eq. (11). Consider a rectangular perturbation of length $2u_0$ centered at $u=0$. That is

$$f(u) \equiv 1$$

and

$$\phi(u) = \begin{cases} 1, & -u_0 \leq u \leq u_0 \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

It is evident that

$$\begin{aligned} M_{ff} &= L, \\ M_{\phi f} &= M_{f\phi} = \min(L, 2u_0+1). \end{aligned} \tag{14}$$

For simplicity we assume $u=0$ is the center of the interval X_j. Thus, Eq. (14) leads to

$$\Delta(L) = \sigma^{-2}\gamma^2\{\min(L, 2u_0+1) - \frac{1}{L}\min[L^2, (2u_0+1)^2]\}.$$

If $L \leq 2u_0+1$, then $\Delta(L)=0$. Suppose $L > 2u_0+1$. In this case

$$\Delta(L) = \sigma^{-2}\gamma[2u_0 + 1 - \frac{(2u_o+1)^2}{L}],$$

and one has to maximize with respect to L the function

$$\frac{1}{\sqrt{L-1}}[(2u_0+1) - \frac{(2u_o+1)^2}{L}].$$

Obviously,

$$L^* = [(2u_0+1)\frac{3+\sqrt{9-\frac{4}{2u_0+1}}}{2}].$$

If $2u_0+1 \gg 1$, then the maximum occurs for

$$L^* \approx 3(2u_0+1).$$

Thus the optimum window size for detecting a rectangular perturbation is roughly three times the width of the perturbation.

3.4. Working with Sets of Intervals

Treating each interval X_j separately permits us to use very standard and simple statistical techniques. In fact, that approach applied simultaneously to all intervals X_j reveals some interesting properties of large data sets. Suppose there exists a physically meaningful partition $\{X_j\}_{j=1}^{J}$. In general, the data for almost all intervals are perturbation free and can be described by the regression model Eq. (3) without the second term,

$$y_i = \theta_j^T f(u_{ji}) + \epsilon_i. \tag{15}$$

Occasionally, a perturbation appears and the perturbation term $\gamma_j^T \phi(u_j)$ becomes significant, that is,

$$\max_u |\gamma_j^T \phi(u)| \gg \sigma,$$

where $E(\epsilon_i^2) = \sigma^2$. We are assuming that the size of a perturbation relative to the size of the interval X_j is small. According to the previous section, the parameter estimators $\hat{\theta}_j$ and $\nu^2(\hat{\theta}_j)$ may be used as descriptors of the j-th interval. An unusual value of $\hat{\theta}_j$ and a relatively large value of $\nu^2(\hat{\theta}_j)$ indicates the presence of a perturbation.

Let us consider a segment of data from measurements of liquid water content of the atmosphere. This data was collected under the auspices of the Atmospheric Radiation Measurement (ARM) project [11]. The segment contains 200,000 observations taken at 20 second intervals, which is a little over 46 days of measurements. We compute a least squares fit of the simple model

$$y_i = \theta_{j1} + \theta_{j2} u_{ji} + \epsilon_i,$$

where the u_{ji} are components of the same centered and normalized vector $(1, 2, \ldots, L)^T$ for each segment j, and $(j-1)L + 1 \leq i \leq jL$.

Fig. 1 shows a plot of $\theta_{j1} \times \theta_{j2}$ for 2,000 segments of size 100. This plot, as well as all other plots in this report are produced with Splus [25]. Approximately 200 points are outside the thin dark elliptical region, indicating the 200 intervals that potentially contain perturbations. The scatterplot appears to be a mixture of two distributions: one tight distribution containing roughly 1800 windows without perturbations and a more diffuse

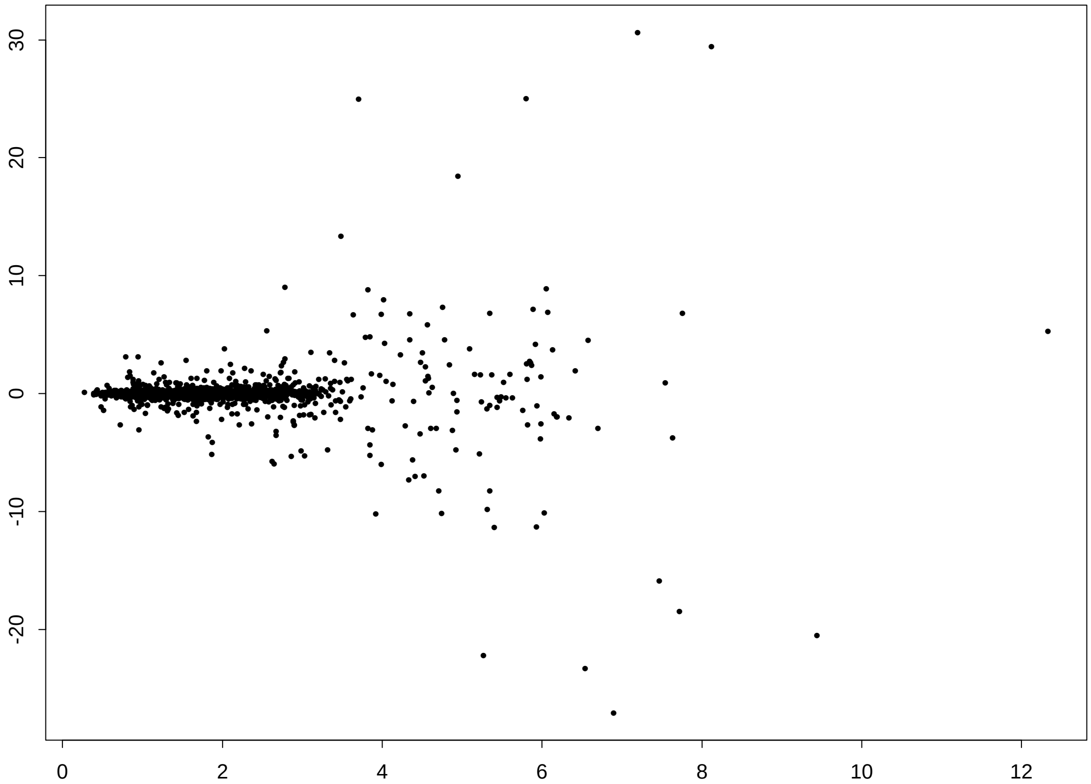

Figure 1: $\hat{\theta}_{j1}$ (horizontal axis) vs. $\hat{\theta}_{j2}$ for liquid water ARM data

distribution containing roughly 200 windows with perturbations. Also note that all points lie in a cone with its vertex at the origin.

A possible data generation model to explain the phenomenon observed in Figure 1 is a piecewise linear data series with a small probability of a single transient perturbation of length 1 in each interval. The location of the transient within an interval follows a uniform distribution and its size is random with a given density. More formally

$$y_i = \theta_{1j} + \theta_{2j} u_{ji} + \pi_j \gamma_j \delta(u_{ji} - u_j) + \epsilon_i, \tag{16}$$

where: $(j-1)L+1 \leq i \leq jL$; $\epsilon_i \sim N(0, \sigma^2)$; π_j are independent Bernoulli random variables such that

$$p(\pi_j = 1) = \alpha, \quad p(\pi_j = 0) = 1 - \alpha,$$

and α is assumed to be small; γ_j is a random variable with density $\psi(\gamma_j)$; $\delta(u_{ji} - u_j) = 1$ if

$u_{ji} = u_j$ and $\delta(u_{ji} - u_j) = 0$ if $u_{ji} \neq u_j$; and u_j are independent, discrete uniform random variables on L points.

We drop the interval subscript j in the following because our model for each interval is the same. Consider the conditional distribution of θ determined by π, γ, and u. If $\pi = 0$, we have the standard case and the conditional distribution of θ is normal, with

$$\begin{aligned} E(\hat{\theta}) &= \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \\ \mathrm{Var}(\hat{\theta}) &= \frac{\sigma^2}{L} \begin{pmatrix} 1 & 0 \\ 0 & \frac{12}{L^2-1} \end{pmatrix}, \end{aligned} \tag{17}$$

where we assume L is odd and $u_{L/2+1} = 0$.

It is convenient to have $\theta = 0$, which we assume in what follows. If $\pi = 1$, then

$$\begin{aligned} \hat{\theta}_1 &= \frac{\gamma}{L} + \frac{1}{L}\sum_i \epsilon_i, \\ \hat{\theta}_2 &= \frac{12}{L(L^2-1)} u\gamma + \frac{12}{L(L^2-1)} \sum_i \epsilon_i u_i, \end{aligned} \tag{18}$$

where $u_i = (2i - L - 1)/2$. Eliminating γ, this gives

$$\begin{aligned} \hat{\theta}_2 &= \frac{12u}{L^2-1}\hat{\theta}_1 - \frac{12u}{L(L^2-1)}\sum_i \epsilon_i + \frac{12}{L(L^2-1)}\sum_i \epsilon_i u_i \\ &= \frac{12u}{L^2-1}\hat{\theta}_1 + \frac{12}{L(L^2-1)}\sum_i \epsilon_i (u_i - u). \end{aligned} \tag{19}$$

Thus, the conditional distribution of $\hat{\theta}_2$ given $\hat{\theta}_1$, $\pi = 1$, and u is the normal distribution with conditional mean

$$E(\hat{\theta}_2 | \hat{\theta}_1, u, \pi = 1) = \frac{12u\hat{\theta}_1}{L^2-1} \tag{20}$$

and variance

$$\mathrm{Var}(\hat{\theta}_2 | \hat{\theta}_1, u, \pi = 1) = \frac{12\sigma^2}{L(L^2-1)}(1 + \frac{12u^2}{L^2-1}). \tag{21}$$

Recalling that $u_i = -u_{L-i+1}$, we see that the conditional distribution of $\hat{\theta}_2$ given $\hat{\theta}_1$, u, and $\pi = 1$ is symmetric with respect to zero. The conditional mean Eq. (20) of $\hat{\theta}_2$ depends on γ through $\hat{\theta}_1$ and the variance Eq. (21) is independent of γ or $\hat{\theta}_1$. For large values of γ we have large values of $\hat{\theta}_2$ and the constant variance becomes relatively negligible. The intervals

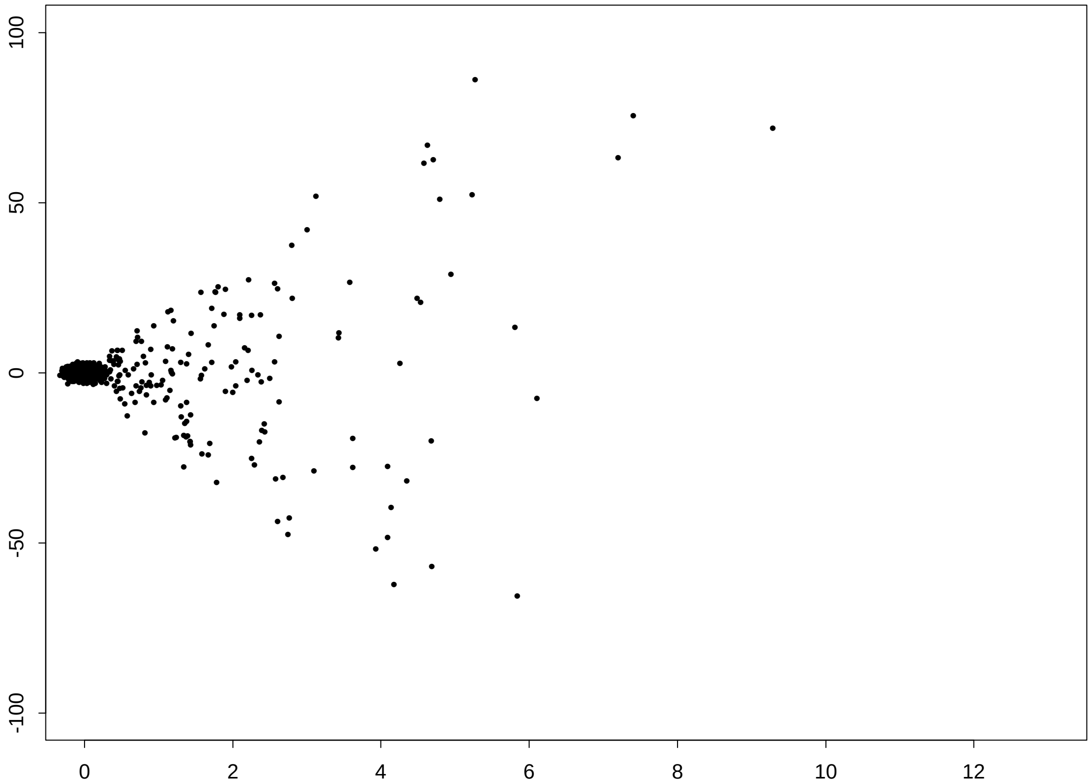

Figure 2: $\hat{\theta}_1$ vs. $\hat{\theta}_2$ for generated data

with perturbations are scattered around rays emanating from the origin and described by equation Eq. (20). Each ray represents a different location of the transient (determined by u_l) within a window. For a given ray direction, γ controls the distance from the origin. Together, these rays result in a cone shaped region much like that observed in the ARM data.

We generated 100,000 observations with model Eq. (16) and parameters $\theta_1 = \theta_2 = 0$, $\alpha = .1$, $\sigma = 1$, and $\psi(\gamma) \sim \exp(.01)$. The results of fitting a simple linear model to partitions of size 50 (analogous to Fig. 1 for the ARM data) are displayed in Fig. 2. Here one can see similarity to Fig. 1 and that the distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$ are of the types described by Eqs. (17), (19), respectively, for the weights $1 - \alpha$ and α.

3.5. Use of Local Regression in Overlapping Intervals

To this point we have only considered analyses performed on mutually exclusive intervals, as described in Eq. (1). In essence, we have suggested partitioning the bivariate data set into contiguous subsets and analyzing each as a separate problem. While computationally convenient, such partitioning of the data can in some cases lead to reduced power to detect perturbations, particularly in cases where the partitioning is arbitrary or where the appearance of perturbation can occur anywhere with respect to the partitioning.

To demonstrate this, consider a modification to the model perturbation Eq. (13) used for the purpose of demonstrating an optimal analysis of interval length, L^* Eq. (11), in Section 3.3. Assuming for simplicity that L is even and that $x_i = i$, the analysis is based on the scaled location or time variable

$$-\frac{L}{2} \leq u \leq \frac{L}{2}.$$

Now, in place of Eq. (13), let the perturbation be defined more generally as:

$$\phi(u) = \begin{cases} 1, & u_* \leq u \leq u^* \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

where $u^* - u_* + 1 = P$, the length of the perturbed subinterval. However, in this more general case, the subinterval affected by the data is no longer necessarily "centered" with respect to, or even contained entirely within, the analysis interval. Δ is now determined by the size of the subset of perturbed data which lies within the analysis interval:

$$P_L = \min(u^*, \frac{L}{2}) - \max(u_*, -\frac{L}{2})$$

and

$$\Delta = \sigma^2 \gamma (P_L - \frac{P_L}{L}).$$

Hence, if $P \ll L$, Δ can be reduced by approximately a factor of 2 if half the data affected by the perturbation are not included in the analysis.

One way to avoid this difficulty is to analyze overlapping subsets, or a "sliding window", of the data. More specifically, substitute

$$\begin{aligned} Y_j &= \{y_i\}_{i=j}^{j+L-1}, \\ X_j &= \{x_i\}_{i=j}^{j+L-1}. \end{aligned}$$

for their counterparts in Eq. (1). This will certainly require more individual analyses; for windows of length L, $N - L + 1$ separate analyses will be needed, rather than the $\frac{N}{L}$ analyses suggested in the previous section. However, if the individual analyses are relatively computationally fast, and the user is only interested in identification and characterization of the perturbed subsets (rather than saving summaries of all subsets), then ideas similar to those discussed above but based on a sliding window approach may be more satisfactory. Continuing with the modification of the example perturbation Eq. (13) described by Eq. (22), sensitivity to the change is no longer dependent upon placement of the initial window, since every interval of a given length is examined. Specifically, windows for which Δ is maximized for given L, i.e. those for which the entire perturbed subsequence is included, are now examined.

As for computational effort, we will note that the basic matrices used in regression analysis can be modified via rank-two updating, rather than recomputed from scratch in this setting. For example,

$$M_{jff} = M_{j-1ff} - f(u_{j-11})f(u_{j-11})^T + f(u_{jL})f(u_{jL})^T,$$

and similarly for $M_{jf\phi}$. Rank-two updating is also possible for M_{ff}, saving explicit inversion or solution of its affiliated equation at each step, but this may not be a serious component of the computation unless the dimension of $f()$ is larger than we have indicated here.

Of course, updating of these matrices will not be necessary if values of x are equally spaced in the series, since the moment matrices will be the same for each interval, as previously noted in Section 3.2. In this case, we can simplify expression of the regression coefficient estimates to

$$\hat{\theta}_j = QY_j,$$

where the matrix Q is the same for each interval. Hence, even if the ϵ_i are uncorrelated as assumed above, estimates of $\hat{\theta}$ for windows containing common data will be correlated:

$$Cov(\hat{\theta}_j, \hat{\theta}_{j+l}) = QCov(Y_j, Y_{j+l})Q^T. \tag{23}$$

For $l \geq L$, the central matrix on the right side of Eq. (23) is zero since Y_j and Y_{j+l} contain no common elements. For $l < L$, entries of this matrix for which row and column numbers differ by l are non-zero, and have common value σ^2. This allows us to rewrite covariances explicitly in terms of the elements of Q:

$$Cov(\hat{\theta}_j, \hat{\theta}_{j+l})_{rs} = \sigma^2 \sum_{m=1}^{L-l} (q_{r,m} q_{s,m+l} + q_{s,m} q_{r,m+l}).$$

In particular, for a specific element (r-th) of the estimate vector, the lag-l covariance is

$$2\sigma^2 \sum_{m=1}^{L-l} q_{r,m} q_{r,m+l}.$$

Perhaps more important, single short perturbations will now affect regression estimates in several subsets, resulting in a "trajectory" or path in plots of $\hat{\theta}_j$. Consider again the model Eq. (16) introduced to help explain ARM data analysis results in Section 3.4. We use that model to generate several data sets to illustrate the effect of using a sliding window. Suppose a single perturbation of magnitude γ appears at one point in the series (i.e. $\pi = 1$ only once), where the expectation of y is otherwise zero (i.e. $\theta_1 = \theta_2 = 0$). Expressions for the coefficient estimates are given in Eq. (18), but now the single perturbation will affect analyses in L separate intervals, where the value of u_i will change by one unit in each analysis. Fig. 3 displays the behavior of $(\hat{\theta}_1, \hat{\theta}_2)$ for a simulation of this situation for $N = 101$, $L = 21$, $\gamma = 30$, and $\sigma^2 = 0$; Fig. 4 is similar but with $\sigma^2 = 1$. The 21 points near the top in each figure represent the intervals that contain the perturbation. Figs. 5 and 6 present a slightly more general example, in which a number of consecutive data values are perturbed, in this case $P = 15$. Again $L = 21$, but here $\gamma = 2$, resulting in "integrated" contamination equal to the previous case. Fig. 5 presents the noiseless case ($\sigma^2 = 0$), while Fig. 6 simulates normal errors with $\sigma^2 = 1$. Clearly, different patterns of contamination would produce different traces on these and similar plots. Traces such as those given in the figures may in some cases be helpful in classifying perturbations, either by inspection or by further automated processing.

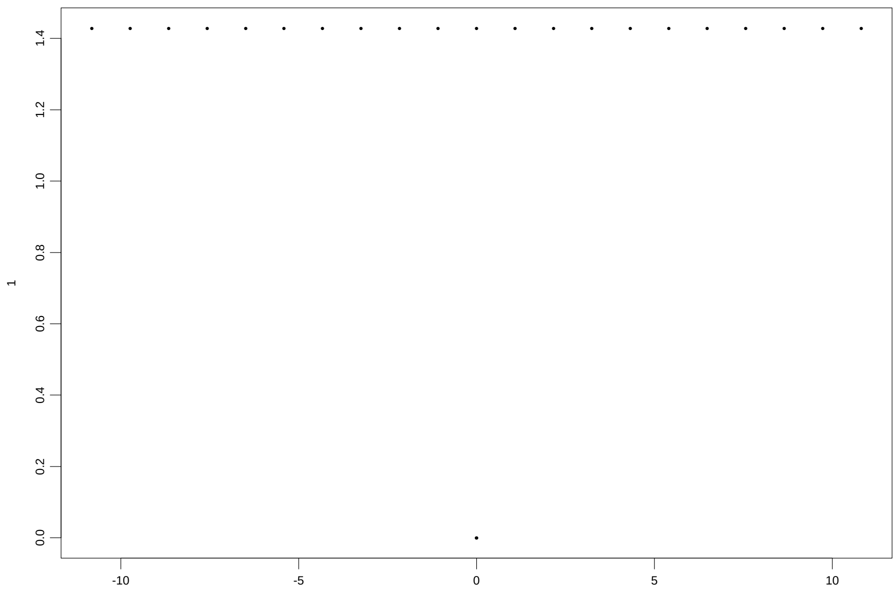

Figure 3: $\hat{\theta}_2$ (horizontal axis) vs. $\hat{\theta}_1$ for $L = 21$ from generated data with $\sigma^2 = 0$ and a single $\gamma = 30$.

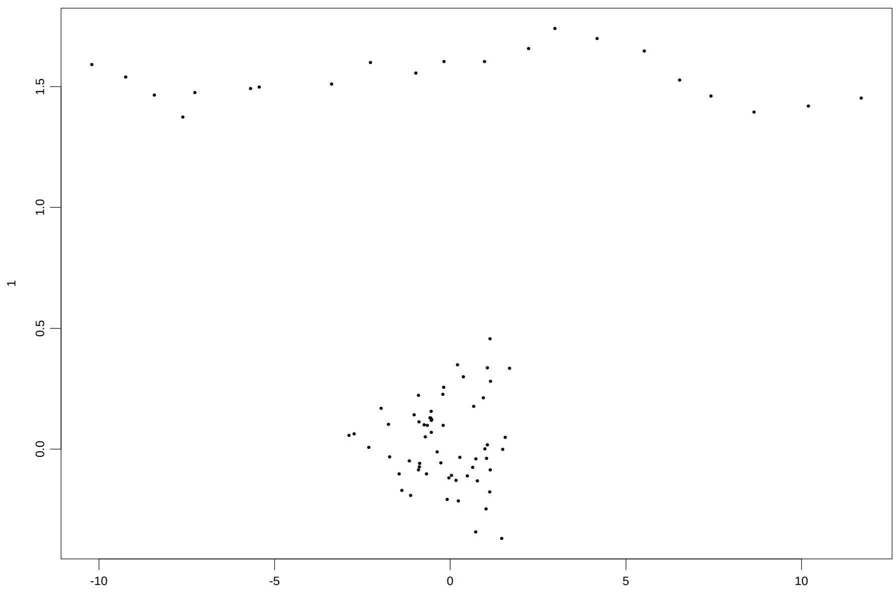

Figure 4: $\hat{\theta}_2$ (horizontal axis) vs. $\hat{\theta}_1$ for $L = 21$ from generated data with $\sigma^2 = 1$ and a single $\gamma = 30$.

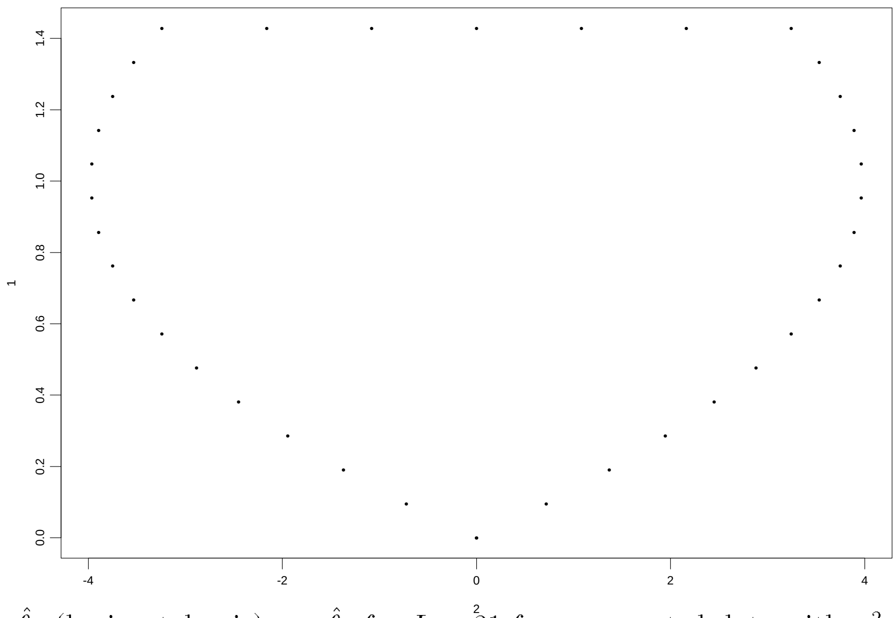

Figure 5: $\hat{\theta}_2$ (horizontal axis) vs. $\hat{\theta}_1$ for $L = 21$ from generated data with $\sigma^2 = 0$ and 15 consecutive $\gamma = 2$.

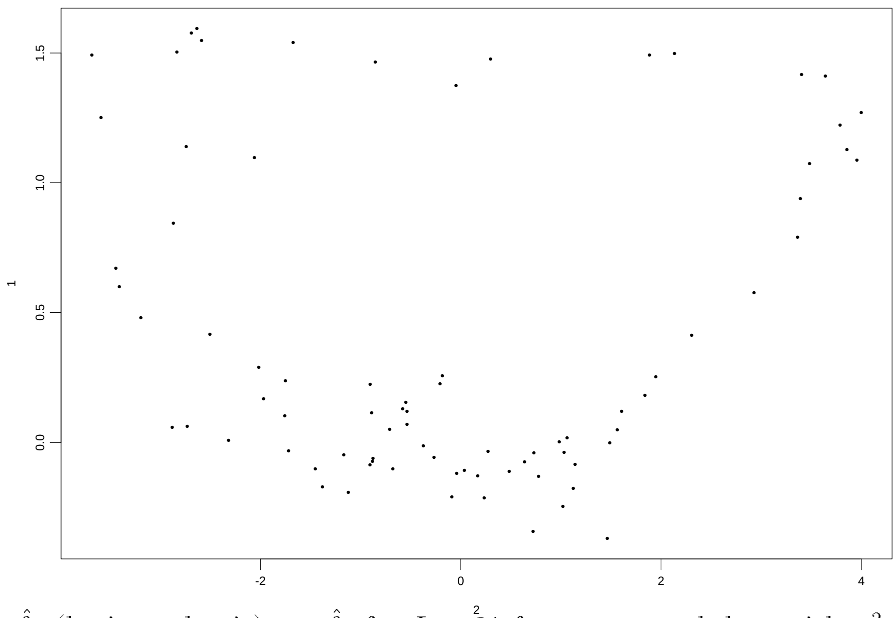

Figure 6: $\hat{\theta}_2$ (horizontal axis) vs. $\hat{\theta}_1$ for $L = 21$ from generated data with $\sigma^2 = 1$ and 15 consecutive $\gamma = 2$.

4. Regression Models with Random Parameters

The regression models introduced in Section 3.2 do not assume any connection between various intervals involved in the analyses. It appears that we can improve matters by applying some realistic models that describe the relationship between intervals and their descriptors. For now, we exclude the perturbation term $\gamma^T\phi(u)$ from consideration.

4.1. Independent sampling

One of the models with a very "loose" link between intervals is the so called regression model with random parameters (or coefficients):

$$y_{jl} = \theta_j^T f(u_{jl}) + \epsilon_{jl}, \quad j = 1, \ldots, J, \quad l = 1, \ldots, L. \tag{24}$$

We added the subscript j to the terms y and ϵ to emphasize their connection with a particular interval. We make the following assumptions:

1. The random variables ϵ_{jl} are independent (for all j and l) and normally distributed, with $\mathrm{E}(\epsilon_{jl}) = 0$ and $\mathrm{E}(\epsilon_{jl}^2) = \sigma^2$.

2. The parameters θ_j are sampled from a normal distribution, with $\mathrm{E}(\theta_j) = \theta_0$ and $\mathrm{Var}(\theta_j) = D_0$ (or θ_j are independent for various j).

There exist several levels of prior knowledge about the distribution of θ_j.

1. The most important situation is probably the case when the parameters θ_0, the matrix D_0, and the variance σ^2 are known and we want to estimate individual parameters θ_j.

2. In other situations, both parameters θ_0 and interval parameters θ_j are to be estimated, and D_0 and σ^2 are known.

3. Finally, there is the case when all parameters θ_j, θ_0, D, and σ^2 are unknown.

Table 1, adapted from Fedorov, et al. [13], (also see Ch. 4 in Rao [24]) summarizes the estimators in three columns corresponding to the available knowledge. In this table

$$M = \sum_l f(u_{jl}) f^T(u_{jl}),$$

Table 1: Estimators $\hat{\theta}_j$ and $\hat{\theta}_0$ together with their covariance matrices $D_{\hat{\theta}_j}$ and $D_{\hat{\theta}_0}$, respectively, and estimators $\hat{D}_0$ and σ^2 for three situations of prior knowledge.

	Known: θ_0, σ^2, D_0 Unknown: θ_j	Known: σ^2, D_0 Unknown: θ_0, θ_j	Known: Unknown: θ_0, σ^2, D_0, θ_j
$\hat{\theta}_j$	$(D_0^{-1}+\sigma^{-2}M)^{-1}$ $\times(D_0^{-1}\theta_0+\sigma^{-2}y_j)$	$M^{-1}y_j$	$M^{-1}y_j$
$D_{\hat{\theta}_j}$	$(D_0^{-1}+\sigma^{-2}M)^{-1}$	$\sigma^2 M^{-1}$	$\hat{\sigma}^2 M^{-1}$
$\hat{\theta}_0$	–	$J^{-1}M^{-1}\sum_{j=1}^{J} y_j$	$J^{-1}M^{-1}\sum_{j=1}^{J} y_j$
$D_{\hat{\theta}_0}$	–	$J^{-1}(D_0+\sigma^{-2}M^{-1})$	$J^{-1}(\hat{D}_0+\hat{\sigma}^{-2}M^{-1})$
$\hat{\sigma}^2$	–	–	$J^{-1}(L-p)^{-1}\sum_{j=1}^{J}\sum_{l=1}^{L}[y_{lj}-\hat{\theta}_j f(x_l)]^2$
$\hat{D}_0$	–	–	$(J-1)^{-1}\sum_{j=1}^{J}(\hat{\theta}_j-\hat{\theta}_0)(\hat{\theta}_j-\hat{\theta}_0)^T$ $-\hat{\sigma}^{-2}M^{-1}$

$$y_j = \sum_l f(u_{jl})y_{jl}.$$

For the sake of simplicity, given l, u_{jl} are assumed identical for all j.

All calculations using the formulae reported in Table 1 are based on Eq. (4) and, therefore, may be considered technically simple. However, there exists one barrier to overcome, namely, we have assumed no perturbation. We can proceed with perturbations as follows: If their relative frequency is low, then we can use the statistical idea about outliers to modify the estimators from Table 1 (see [17]). For instance, only intervals with

$$\hat{t}_j = (\hat{\theta}_j - \hat{\theta}_0)^T \hat{D}_0 (\hat{\theta}_j - \hat{\theta}_0) \leq t,$$

where the cutoff value t is usually selected a few times larger than the dimension p of θ_j, are included in a second round of calculations.

To conclude this subsection we note that $\hat{\theta}_j$, $\nu^2(\hat{\theta}_j)$, and $\hat{t}_j$ may be used as descriptors for the j-th interval. The whole data set is now characterized by $\hat{\theta}_0$ and $\hat{D}_0$.

4.2. Autoregressive Correlation

Stronger dependencies between various intervals may be easily introduced. For instance, an autoregressive model may be used to connect neighboring θ_j and θ_{j+1} by

$$\theta_{j+1} = A\theta_j + \xi_j, \tag{25}$$

where A is a $p \times p$ matrix and ξ_j is a random vector with $\mathrm{E}(\xi_j) = 0$ and $\mathrm{E}(\xi_j \xi_j^T) = \Xi$. The combined likelihood function for Eqs. (24), (25) may be constructed and minimized with respect to θ_j, A, σ^2, and Ξ. This can lead to a rather complicated computing procedure, even for a moderate size data set. In our case, where data sets are very large, one may proceed with the simplified interval-wise procedure:

Stage 1. Produce the usual least squares estimates of $\hat{\theta}_j$ and σ_j^2 for each j.

Stage 2. Proceed with the minimization of $\left|\sum_j (\hat{\theta}_{j+1} - A\hat{\theta}_j)(\hat{\theta}_{j+1} - A\hat{\theta}_j)^T\right|$ with respect to A. The corresponding $\hat{A}$ is the estimate of A.

The properties of estimates generated by this procedure are well known (see, for instance [3]). In particular, one can analyze the residuals

$$\hat{\theta}_j - \hat{A}\hat{\theta}_j$$

to identify changes in long term trends (see [5]).

4.3. Smoothed Trends

Another way to implement some links between intervals is to introduce deterministic constraints on $\{\theta_j\}_1^J$. For instance, we may be interested in imposing a continuity constraint on the piecewise fitted response. For instance, we can apply the constraint

$$\theta_j^T f(u_{j(L+1)}) = \theta_{j+1}^T f(u_{(j+1)1}), \tag{26}$$

or, for a smoother approximation, we can use

$$\theta_j^T \dot{f}(u_{j(L+1)}) = \theta_{j+1}^T \dot{f}(u_{(j+1)1}),$$

where

$$\dot{f}(u) = \frac{\partial f}{\partial u}.$$

For estimation constrained by Eq. (26), we have to find

$$\{\theta_j\}_1^J = \arg \min_{\{\theta_j\}_1^J} \sum_{j=1}^{J} \sum_{l=1}^{L} \left(y_{jl} - \theta_j^T f(u_{jl}) \right)^2,$$

where $\{\theta_j\}_1^J$ is subject to the condition Eq. (26). The computation of $\{\theta_j\}_1^J$ for large J can be a serious problem. A stepwise procedure may be devised that yields to a reasonable approximation, but the corresponding theoretical analysis is beyond the scope of this project.

5. Basis Selection

In the previous section, the data partition, $\{X_j\}_{j=1}^J$, and the basis function, $f(u)$, were assumed to be given. In the following, we consider some recommendations on the selection of the basis function $f(u)$. As we have to this point, in the following we rely heavily on statistical approaches and, in particular, on regression models of type Eq. (24).

5.1. Known Covariance Kernel

We assume that each set X_j is sufficiently large to be treated as a longitudinal or temporal series with some covariance structure. For each interval X_j it is reasonable to consider

$$Cov(y_{jl_1}, y_{jl_2}) = K(u_{jl_1}, u_{jl_2}) = E\left[\left(y_{jl_1} - E(y_{jl_1}) \right) \left(y_{jl_2} - E(y_{jl_2}) \right) \right].$$

For the sake of simplicity, we assume

$$E(y_{jl}) = 0.$$

The generalization is straightforward for nonzero deterministic trends.

Let $K(u, w)$ be a reproducing, nonnegative definite kernel in L_2, so that

$$(K\phi, \phi) = \int K(u, w)\phi(u)\phi(w)du dw \geq 0$$

for any $\phi(u) \in L_2$. We also assume that the data on X_j is generated by a process that can be represented as

$$y(u) = \int K(u,w)\phi(w)dw, \tag{27}$$

where $\phi(w) \in L_2$.

Let $\{\psi_\alpha(u)\}$ and $\{\lambda_\alpha\}$ be corresponding sets of eigenfunctions and eigenvalues of the kernel K, so that

$$\lambda_\alpha \psi_\alpha(u) = \int K(u,w)\psi_\alpha(w)dw, \quad \lambda_1 \geq \lambda_2 \geq \cdots$$

The set $\{\psi_\alpha(u)\}$ can be orthonormalized, and thus we assume

$$(\psi_\alpha, \psi_\beta) = \int \psi_\alpha(u)\psi_\beta(u)du = \delta_{\alpha\beta}.$$

It is known (see, for example, Ch. 7 in [19]) that

$$\sum_\alpha \lambda_\alpha^2 \leq \int |K(u,w)|^2 dudw < \infty, \tag{28}$$

which also implies $\lambda_\alpha^2 \to 0$. Nonnegative definiteness implies

$$\lambda_\alpha \geq 0, \quad \forall \alpha.$$

There are two results from integral equation theory that provide a foundation for our following considerations. The first result, usually referred to as Mercer's theorem, states that the kernel K may be expanded as

$$K(u,w) = \sum_{\alpha=1}^{\infty} \lambda_\alpha \psi_\alpha(u)\psi_\alpha(w),$$

the series being uniformly and absolutely convergent. The second result, known as the Hilbert-Schmidt theorem, which holds for any function satisfying Eq. (27), is

$$y(u) = \sum_{\alpha=1}^{\infty} \theta_\alpha \psi_\alpha(u),$$

where

$$\theta_\alpha = \int y(u)\psi_\alpha(u)du = (y, \psi_\alpha).$$

Also,

$$\theta_\alpha = h_\alpha \lambda_\alpha,$$

where

$$h_\alpha = \int \phi(u)\psi_\alpha(u)du = (\phi, \psi_\alpha).$$

From Bessel's inequality it follows that

$$\sum_{\alpha=1}^{\infty} h_\alpha^2 \leq \int \phi^2(u)du < \infty, \tag{29}$$

which implies $h_\alpha^2 \to 0$ faster than $\alpha^{-1} \to 0$. Subsequently,

$$\lim_{\alpha\to\infty} \theta_\alpha^2 = 0. \tag{30}$$

Moreover, from Eqs. (28),(29) it follows immediately that

$$\lim_{\alpha\to\infty} \alpha^2\theta_\alpha^2 = 0, \tag{31}$$

so that θ_α^2 goes to zero faster than α^{-2}.

It is known that

$$||y(u) - y_p(u)|| \leq \lambda_{p+1}^2 ||\phi(u)||,$$

where $\phi(u)$ is defined by Eq. (27) and

$$y_p(u) = \sum_{\alpha=1}^{p} \theta_\alpha \psi_\alpha(u).$$

Next, we take a step toward developing the approach described in Section 4. Assume for all intervals X_j that

$$y(u) = \sum_{\alpha=1}^{\infty} \theta_{\alpha j}\psi_\alpha(u), \tag{32}$$

where $u = x - x_{j_0}$, $\theta_{\alpha j}$ are random with

$$E(\theta_{\alpha j}) = 0, \quad E(\theta_{\alpha j}\theta_{\beta j}) = \lambda_\alpha \delta_{\alpha\beta},$$

for α, $\beta = 1, 2, \ldots$. Note, there are no observational errors in the $y(u)$ representation Eq. (32). Thus, $y(u)$ is random only if one or more of the coefficients $\{\theta_\alpha\}$ is random.

It is straightforward to show

$$E[y(u_1)y(u_2)] = K(u_1, u_2).$$

Assuming the intervals X_j are "dense" enough to neglect the discreteness of $u_{l_j} = x_l - x_{j_0}$, we can use the results corresponding to Eqs. (6)-(10). In particular, it follows that the parameter

$$\theta_{\alpha j} = \int y(u_j)\psi_\alpha(u_j)du_j$$

may be used as a descriptor for the j-th interval. Of course, we do not gain in data reduction if all the parameters θ_{α_j} are saved, but it is evident from the limits Eqs. (30),(31) that even a few of the lowest order parameters can be expected to provide a good description of the considered interval. We note the condition $\theta_\alpha^2 \gg \lambda_\alpha$ indicates the presence of a perturbation.

From the discussion above, it is clear that the use of the eigenfunctions $\{\psi_\alpha\}$ is convenient and, also, guarantees the difference between the observed $y_j(u)$ and the finite expansion $\sum_{\alpha=1}^{p} \theta_{\alpha j}\psi_\alpha(u)$ will be "small" for sufficiently large p. However, the question is open whether the choice of basis functions $\{\psi_\alpha\}$ is optimal in some reasonable sense.

To address the optimality question, consider the function

$$\begin{aligned}\Phi(K, \{f_\alpha\}_1^p) &= E[\min_\gamma ||y(u) - \sum_{\alpha=1}^{p} \gamma_\alpha f_\alpha(u)||^2] \\ &= E[\min_\gamma \int (y(u) - \sum_{\alpha=1}^{p} \gamma_\alpha f_\alpha(u))^2 du].\end{aligned}$$

It can be shown that the basis

$$\{\psi_\alpha(u)\}_1^p = \arg\min \Phi(K, \{f_\alpha\}_1^p), \tag{33}$$

where minimization is over $\{f_\alpha : (f, f) = 1\}_1^p \subset L_2$, is optimal. The proof of this result follows from the observation that this optimization problem is equivalent to the Rayleigh problem

$$\{\psi_\alpha\}_1^p = \arg\max \sum_{\alpha=1}^{p} (f_\alpha, K f_\alpha).$$

According to Rayleigh's theorem [19], the corresponding maximum is $\sum_{\alpha=1}^{p} \lambda_\alpha$ and $\{\psi_\alpha\}_1^p$ are the first p eigenfunctions of the kernel K. In addition, note that the optimal solution Eq. (33) is not unique. Indeed, we may select any bounded linear transformation of $\{\psi_\alpha(u)\}_1^p$ as a new basis without changing the value of the function Φ.

There are several well known example processes for which analytic expressions for the kernel exist and the corresponding eigenvalues and eigenfunctions are known. For instance, it is known that for Brownian motion the kernel is

$$K(s,t) = \min(s,t), \quad 0 \leq s,t \leq 1$$

and that the eigenvalues and eigenvectors of the kernel are

$$\begin{aligned} \lambda_\alpha &= \frac{1}{(\alpha - \frac{1}{2})^2 \pi^2}, \\ \psi_\alpha(s) &= \sqrt{2} \sin(\alpha - \frac{1}{2}) \pi s. \end{aligned}$$

For the Brownian bridge,

$$K(s,t) = \min(s,t)[1 - \max(s,t)], \quad 0 \leq s,t \leq 1, \tag{34}$$

and we have

$$\begin{aligned} \lambda_\alpha &= \alpha^{-2} \pi^{-2}, \\ \psi_\alpha(s) &= \sqrt{2} \sin(\alpha \pi s). \end{aligned}$$

Assuming

$$\begin{aligned} s = s_j &= \frac{x - x_{j0}}{r_j}, \\ x &\in X_j, \\ r_j &= \max_x X_j - \min_x X_j, \\ x_{j0} &= \min_x X_j, \end{aligned}$$

one can use

$$\hat{\theta}_{j\alpha} = \sqrt{2}\int y(s_j)\sin(\alpha - \frac{1}{2})\pi s_j ds_j$$

for the Brownian process and

$$\hat{\theta}_{j\alpha} = \sqrt{2}\int y(s_j)\sin(\alpha\pi s_j)ds_j$$

for the Brownian bridge as descriptors. Evidently, in both cases $\hat{\theta}^2_{j\alpha}$ diminishes faster than α^{-2}.

Another example, which is a popular choice for modeling correlated observations, is the Poisson kernel

$$\begin{aligned} K(s,t) &= \frac{1-\rho^2}{2}[1 - 2\rho\ \cos[2\pi(s-t)] + \rho^2]^{-1} \\ &= \frac{1}{2} + \sum_{\alpha=1}^{\infty}\rho^{\alpha}\ \cos[2\pi\alpha(s-t)], \end{aligned}$$

where $0 \le s, t \le 1$ and $0 < \rho < 1$. The eigenvalues and eigenfunctions are

$$\lambda_0 = 1, \quad \lambda_{2\alpha-1} = \lambda_{2\alpha} = \rho^{\alpha}, \quad \alpha = 1, 2, \ldots,$$

and

$$\psi_0(s) = 1, \quad \psi_{2\alpha-1}(s) = \sqrt{2}\cos(2\pi\alpha s), \quad \psi_{2\alpha} = \sqrt{2}\sin(2\pi\alpha s),$$

respectively. The shape of the Poisson kernel is easily controlled by the parameter ρ.

5.2. Known Covariance Kernel: Discrete Case

As discussed earlier, for data sets with high density X_j we can use the continuous approximation described in Section 4.3. In cases where that condition does not hold, more direct, and usually more computationally demanding, approaches must be considered.

We assume the covariance structure of y_j on X_j is known and is the same for all j. Let

$$\begin{aligned} K_{l_1,l_2} &= K(u_{l_1 j}, u_{l_2 j}) \\ &= E[(y_{l_1} - E(y_{l_1}))(y_{l_2} - E(y_{l_2})], \end{aligned}$$

where $j = 1, \ldots, \jmath$ and $l_1, l_2 = 1, \ldots, L$. In other words, we assume there are no perturbations or short-term tends. Following the idea of Section 4.3, we introduce the eigenvalues and eigenvectors of the covariance structure $K = \{K_{l_1,l_2}\}_1^L$, so that

$$\begin{aligned} \lambda_\alpha \psi_\alpha &= K\psi_\alpha, \quad \lambda_1 \geq \lambda_2 \geq \ldots \lambda_L \geq 0 \\ \psi_\alpha^T \psi_\beta &= \delta_{\alpha\beta}, \qquad \alpha, \beta = 1, \ldots, L. \end{aligned} \tag{35}$$

Any vector y_j may be represented in the form

$$y_j = \sum_{\alpha=1}^{L} \theta_{\alpha j} \psi_\alpha, \tag{36}$$

where

$$\theta_{\alpha j} = y_j^T \psi_\alpha. \tag{37}$$

Obviously,

$$\mathrm{Var}(y_j) = K,$$

and

$$\mathrm{Var}(\theta) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_L \end{pmatrix}.$$

Again, to simplify the subsequent expressions we assume $E(y_j) = 0$. From the theory of nonnegative definite matrices (see, for example, Ch. 1 in Rao [24]) it is known that Eq. (35) has a solution and that

$$K = \sum_\alpha \lambda_\alpha \psi_\alpha \psi_\alpha^T.$$

From Eq. (36) it follows that

$$\theta_j^T = (\theta_{j1}, \ldots, \theta_{jL})$$

is a perfect descriptor of y_j. Unfortunately, $\mathrm{dimy_j} = \mathrm{dim}\theta_\mathrm{j}$ and, therefore, this descriptor does not reduce the size of the data set being considered. The natural way to reduce size is to use the approximate version of Eq. (36)

$$\tilde{y}_j = \textstyle\sum_{\alpha=1}^{p} \theta_{\alpha j} \psi_\alpha, \quad p < L,$$

where θ_{α_j} is defined by Eq. (37). Of course, it is essential to know the error in the approximation $\tilde{y}_j$ of y_j. To answer that, we introduce some measures of similarity between y_j and $\tilde{y}_j$. One of the possibilities is

$$s^2 = E[(y_j - \tilde{y}_j)^T (y_j - \tilde{y}_j)]. \tag{38}$$

From Eqs. (35) and (37) and the definitions of y_j and $\tilde{y}_j$, it follows immediately that

$$s^2 = \sum_{\alpha=p+1}^{L} \lambda_\alpha.$$

It is useful to note that θ_{α_j} may be calculated not only by Eq. (37), but also as the solution to the least squares problem

$$\{\theta_{\alpha j}\}_1^p = \arg \min_{\{\theta_\alpha\}_1^p} \sum_{l=1}^{L} (y_{jj} - \sum_{\alpha=1}^{p} \theta_\alpha \psi_{\alpha l})^2. \tag{39}$$

We now consider the following problem: Is it possible to find a better set of basis functions than $\{\psi_\alpha\}_1^L$? In compliance with the criterion Eq. (38), we look for

$$\begin{aligned} f^* &= \arg\min_f E[(y_j - \tilde{y}_j)^T (y_j - \tilde{y}_j)] \\ &= \arg\min_f E[\min_\theta \sum_{l=1}^{L} (y_{jl} - \sum_{\alpha=1}^{p} \theta_\alpha f_{\alpha l})^2]. \end{aligned} \tag{40}$$

We assume the basis f is orthonormal,

$$\sum_{l=1}^{L} f_{\alpha l} f_{\beta l} = \delta_{\alpha\beta}.$$

After some manipulation, Eq. (40) may be transformed to a problem that is well studied (see, for example, p. 63 in Rao [24]). The answer is amazingly simple, namely,

$$f_\alpha^* = \psi_\alpha.$$

Moreover,

$$\arg\min_f E[\min_\theta \sum_{l=1}^{L} (y_{l_j} - \sum_{\alpha=1}^{p} \theta_\alpha f_{\alpha l})^2] = \sum_{\alpha=p+1}^{L} \lambda_\alpha.$$

A reader familiar with the method of principal components may find that the results described to this point are mathematically identical to results concerning the optimal properties of principal components. The analogy of the principal components method (see [24], p.593) enables us to describe even stronger properties than Eq. (40) about the optimal choice of a basis. Let

$$f^* = \arg\min_f ||E[(y_j - \tilde{y}_j)(y_j - \tilde{y}_j)^T]||, \tag{41}$$

where $\tilde{y}_j = \theta^T f(u_{l_j})$ and $||A|| = \sqrt{\sum_{l,k} |a_{lk}|^2}$ is the Euclidean norm. The solution of Eq. (41) is again

$$f_\alpha^* = \psi_\alpha, \quad \alpha = 1, \ldots, p.$$

In Eqs. (40) and (41) we have compared y_j and $\tilde{y}_j$ through the average difference between the two vectors. Another way to compare two random vectors is to compare their distributions. By definition, both of those distributions have zero expectation. Thus, we need to examine the difference in the covariance matrices. For instance, we can compare

$$\mathrm{Var}(y_j) \text{ and } \mathrm{Var}(\tilde{y}_j) = \mathrm{Var}(\sum_{\alpha=1}^{p} \gamma_{\alpha j} f_\alpha),$$

where $\{f_\alpha\}_1^p$ is a set of orthonormal vectors and

$$\gamma_{\alpha_j} = y_j^T f_\alpha.$$

We have to find

$$f^* = \arg\min_f ||\mathrm{Var}(y_j) - \mathrm{Var}(\tilde{y}_j)||.$$

Again, the solution is

$$f_\alpha^* = \psi_\alpha,$$

with

$$\begin{aligned} \mathrm{Var}(\tilde{y}_j) &= \sum_{\alpha=1}^{p} \lambda_\alpha \psi_\alpha \psi_\alpha^T, \\ \min_f ||\mathrm{Var}(y_j) - \mathrm{Var}(\tilde{y}_j)|| &= \sqrt{\sum_{\alpha=p}^{L} \lambda_\alpha^2}. \end{aligned}$$

The above results may be derived from this more general result (see [24]). For any matrix B of rank p,

$$\min_B ||A - B|| = \sqrt{\sum_{\alpha=p+1}^{L} \lambda_\alpha^2}$$

and the minimum is attained for

$$B = \sum_{\alpha=1}^{p} \lambda_\alpha \psi_\alpha \psi_\alpha^T,$$

where

$$\lambda_\alpha \psi_\alpha = A\psi_\alpha, \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_L \geq 0.$$

Like in the continuous case, a few values

$$\theta_{\alpha j} = y_j^T \psi_\alpha, \quad \alpha = 1, \ldots, p,$$

are used as descriptors. If

$$\sum_{\alpha=1}^{p} \theta_{\alpha j}^2 \lambda_\alpha^{-1} > \chi_{1-\delta}^2(p),$$

where δ is some small positive real number and t is the corresponding confidence level for the χ^2-distribution with p degrees of freedom, the presence of a perturbation in the j-th interval is quite likely. Similar to the description of α given in Section 3.3, we can say that $\delta \times N$ is approximately the number of falsely stored usual segments.

5.3. Unknown Covariance Kernel

It is convenient to start with the discrete case. Assume that by some rather conservative method, meaning one whereby the probability of identifying an interval with a perturbation is relatively large, we select a sufficiently large number of perturbation free, or "standard", intervals. As we are concerned with large data sets and are working under the assumption that perturbed intervals are the exception rather than the rule, identifying such a collection of standard intervals is not a difficult problem.

Given a collection of standard data, $\{y_{jl}\}_{l=1}^{L}$, we compute the covariance matrix

$$\hat{K}_{l_1 l_2} = \frac{1}{J} \sum_{j=1}^{J} (y_{j,l_1} - \bar{y}_j)(y_{j,l_2} - \bar{y}_j). \tag{42}$$

The discrete basis functions, $\hat{\psi}_\alpha$, are defined as eigenvectors of $\hat{K}$, so that

$$\hat{K}\hat{\psi}_\alpha = \lambda_\alpha \hat{\psi}_\alpha. \tag{43}$$

The inclusion of $\bar{y}_j$ in Eq. (42) is needed to remove the effect of possible long-trends in the data set. We assume $J \gg L$ to avoid some theoretical difficulties related to accuracy of ψ_α estimations. Note that

$$\bar{y}_j = \frac{1}{L}\sum_{l=1}^{L} y_{jl}$$

is not the best choice for estimating $\bar{y}_j$, because it makes rank$\hat{K} = L - 1$ and can lead to difficulties if the interval length L is small.

In the presence of a long-term trend, which we assume can be explained and subsequently modeled, the term $\bar{y}_j$ in Eq. (42) can be replaced by a more sophisticated prognosis. For instance, we may use

$$\hat{y}_{ji} = \hat{\beta}^T r(x_{ji}), \tag{44}$$

where

$$\begin{aligned} \hat{\beta} &= M^{-1}Y, \\ M &= \sum_{ij} r(x_{ji}) r^T(x_{ji}), \\ Y &= \sum_{ij} y_{ji} r(x_{ji}), \end{aligned}$$

$r(x)$ is a vector of given functions, and the summation with respect to the index j is over a few neighboring intervals. Theoretically, Eq. (44) is not the best choice for estimating $\bar{y}_j$, the most obvious reason being that correlations between neighboring observations are neglected. More sophisticated procedures, such as those due to Dempster, et al. [8], and Fedorov, et al. [13], can be developed. However, those procedures prove beneficial only when the total number of observations is small, which is not the case being investigated in this study. Recall that accuracy in the usual sense of modeling, in this instance prediction, is not of first order importance. We only require a reasonable, but fast, estimator of the trend, because our objective is to recognize perturbations rather than to accurately model them.

In the continuous case, the first step is to select an appropriate analytic form for the kernel $K(u_1, u_2)$. Usually that form includes unknown parameters. For example, for kernels of the Brownian bridge type Eq. (34), we can consider the three parameter family

$$\begin{aligned} K(u_1, u_2) &= K(u_1, u_2; \sigma^2, a, b), \\ &= \sigma^2 \min(u_1, u_2)[b - \max(u_1, u_2)]. \end{aligned}$$

where $a \leq u_1, u_2 \leq b$. Then, estimates of the parameters result from the optimization

$$(\sigma^2, \hat{a}, \hat{b}) = \min_{(\sigma^2, a, b)} \sum_{l_1 l_2} (\hat{K}_{l_1 l_2} - K(u_1, u_2; \sigma^2, a, b))^2. \tag{45}$$

From a statistical point of view, the estimator defined by Eq. (45) is not the best, but in the case that the number of standard, or usual, intervals is large, the theoretical shortcomings of Eq. (45) are not important. Given values for the parameters $(\sigma^2, \hat{a}, \hat{b})$, eigenfunctions for the kernel can be found numerically using standard numerical linear algebra software.

5.4. Moving Windows and Unknown Covariance Kernels

In this section we describe a specific model that incorporates several ideas discussed earlier, including sliding, or moving windows Eq. (2) and unknown covariance kernels Eq. (42). We then discuss how this method is motivated by a point of view different from the standard statistical approach.

Let $\{y_i\}_{i=1}^{N+n}$, where $n \ll N$, denote a given time series. Given a value for n, consider the vector

$$\mathbf{y}_i = (y_i, y_{i+1}, y_{i+2}, \ldots, y_{i+(n-1)})^T, \quad i = 1, 2, \ldots, N. \tag{46}$$

Let the matrix Y be defined by $Y_i = \mathbf{y}_i$, that is, the i-th column of Y is $\mathbf{y}_i$. Then, comparing with $\hat{K}$ Eq. (42), define the matrix

$$K = \frac{1}{N} Y Y^T = \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i \mathbf{y}_i^T.$$

The matrix K is symmetric, nonnegative. We denote the eigenvalue, eigenvector pairs of K by $\{(\lambda_\alpha, \psi_\alpha)\}_{\alpha=1}^{n}$, where

$$\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n. \tag{47}$$

Define the matrix Ψ by

$$\Psi = (\psi_1, \psi_2, \ldots, \psi_n),$$

meaning that the j-th column of Ψ is the eigenvector ψ_j. Next, define the orthonormal transformation

$$\theta = \Psi(\mathbf{y}) = \Psi^T \mathbf{y}. \tag{48}$$

For nonsingular Ψ, the vector θ_i contains all the information about the vector $\mathbf{y}_i$; that is,

$$\mathbf{y}_i = \Psi\theta_i = \sum_{\alpha=1}^{n} \theta_{i\alpha}\psi_\alpha.$$

Also, note that the optimization

$$\theta_i = \arg\min_\theta \sum_{l=1}^{n}(y_{il} - \sum_{\alpha=1}^{n} \theta_\alpha \psi_{\alpha l})^2$$

can be interpreted as parameter fitting for the linear model

$$y_{il} = \sum_{\alpha=1}^{n} \theta_{i\alpha}\psi_{\alpha l}. \tag{49}$$

For most practical situations, the eigenvalues λ_α decrease rapidly with increasing α. Therefore, the regression model Eq. (49) can be replaced by

$$y_{il} = \sum_{\alpha=1}^{m} \theta_{i\alpha}\psi_{\alpha l} + \epsilon_{il}, \tag{50}$$

where m identifies the range of eigenvalues considered to be significant. The remainder ϵ_{il} is equivalent to "noise" in the statistical approach. Referring to comments following Eq. (40), it follows that

$$\frac{1}{N}\sum_{i=1}^{N} \min_\theta \sum_{l=1}^{n}(y_{il} - \sum_{\alpha=1}^{m} \theta_{i\alpha}\psi_{\alpha l})^2 = \sum_{\alpha=m+1}^{n} \lambda_\alpha.$$

Note that all the preceding formulas can be generalized for the "lagged" case where the definition of the vector $\mathbf{y}_i$ Eq. (46) is replaced by

$$\mathbf{y}_i = (y_i, y_{i+k}, y_{i+2k}, \ldots, y_{i+(n-1)k})^T. \tag{51}$$

The modeling approach described here and applied below to the perturbation detection problem, while in perfect keeping with regression modeling concepts as described in Section 3, is motivated by the theory of nonlinear dynamical processes. An excellent review of the ergodic theory of dynamical systems is given by Eckmann and Ruelle [12], while descriptions of the application of that theory to time series analysis are provided by Abarbanel, Brown, Sidorowich, and Tsimring[1], and Grassberger, Schreiber, and Schaffrath[15]. Tong [30] provides a bridge between the mainstream statistical approach and the dynamical systems point of view. We are going to describe the dynamical systems point of view.

Using $E^{\bar{n}}$ to denote $\bar{n}$-dimensional Euclidean space, let

$$f : E^{\bar{n}} \to E^{\bar{n}}$$

be a diffeomorphism. Suppose $M \subset E^{\bar{n}}$ is a compact, $\bar{n}$-dimensional differentiable manifold and that f restricted to M is a diffeomorphism of M. Let A be a compact subset of M such that f maps A onto A. Further, let $U \supset A$ be an open subset of M such that

$$\lim_{i \to \infty} f^i(U) = A.$$

Finally, let p be an ergodic probability measure on A with respect to the transformation f, which is to say that, if $V \subset A$ is measurable with respect to p, then $p(f(V)) = p(V)$.

We refer to $E^{\bar{n}}$ as "state space" and to A as an "attractor." The point of view taken here is that nonlinear dynamical processes, including chaotic systems, are represented by triples of the form (A, f, p), that is, an attractor in state space, a diffeomorphic mapping of that attractor onto itself, and an ergodic probability measure with respect to that attractor and diffeomorphic mapping.

Suppose $a \in A$ is the state of the experimental system at the instant we begin to observe it. Then, $\{f^i(a)\}_{i=0}^{\infty}$ is a time series of states visited by the process. Now, suppose $y : A \to \mathbb{R}$ is a real-valued function, or observable, on A. Then,

$$y_i = y(a_i) = y(f^i(a)), \quad i = 0, 1, 2, \ldots,$$

is a real-valued time series of the experimental system.

The following is a theorem due to Takens[29] that constitutes the foundation for the time series analysis methods described here. As used in the statement of the theorem, "smooth" means at least C^2.

Theorem 1. *(Takens) Let $M \subset E^{\bar{n}}$ be a compact, differentiable manifold of dimension $\bar{n}$. Further, let $f : M \to M$ be a smooth diffeomorphism of M and let $y : M \to I\!R$ be a smooth, real-valued function on M. It is a generic property that the map*

$$\Phi : M \to E^{2\bar{n}+1}$$

defined by

$$\Phi(a) = (y(a), y(f^1(a)), \ldots, y(f^{2\bar{n}}(a))^T, \quad a \in M,$$

where T is matrix transpose, is a smooth embedding.

Consider the object $\tilde{A}_y \subset E_y^n$, defined as $\tilde{A}_y = \{\mathbf{y}_i\}_{i=0}^{\infty}$ with $\mathbf{y}_i$ defined by Eq. (51). Provided $n \geq 2\bar{n} + 1$, Takens' reconstruction theorem implies the mapping

$$\tilde{\Phi} : A \to \tilde{A}_y$$

defined by

$$\tilde{\Phi}(a_i) = \mathbf{y}_i, \quad i = 0, 1, 2, \ldots,$$

produces a smooth embedding $\tilde{\Phi} : A \to E_y^n$. Define

$$\begin{aligned} \tilde{f}_y &= \tilde{\Phi} \circ f \circ \tilde{\Phi}^{-1} \\ \tilde{p}_y &= p \circ \tilde{\Phi}^{-1}. \end{aligned}$$

Then, the triple $(\tilde{A}_y, \tilde{f}_y, \tilde{p}_y)$ constitutes a faithful representation of the dynamical system (A, f, p).

At this point Broomhead and King [7] introduce the orthonormal transformation

$$\Psi : E_y^n \to E_\theta^n$$

defined by Eq. (48) as a tool for optimally choosing values for the parameter pair (n, k). Lawkins, et al. [21] discuss that process in a general setting, while Downing, et al. [9] discuss it in the context of the perturbation identification problem. In any case, from the standpoint of the theory of nonlinear dynamical systems, there results the mapping

$$A \xrightarrow{\tilde{\Phi}} \tilde{A}_y \xrightarrow{\tilde{\Psi}} \tilde{A}_\theta$$

and the representation $(\tilde{A}_\theta, \tilde{f}_\theta, \tilde{p}_\theta)$ in E^n_θ of the process.

6. Applications of Perturbation Detection

In this section we consider the application of four distinct detection techniques to two time series representing very different processes, one being atmospheric and the other neurophysiological. One of the analysis techniques is regression, which is discussed in Section 3. Some discussion of the other three techniques is given here, but references are provided in lieu of complete descriptions. The primary purpose of this section is to present analysis results in a self consistent fashion so that a logical comparison of the four techniques can be made. Finally, we note that the data used here for demonstration purposes was selected because they are relatively rich in features, and so these data purposely do not represent the field condition where we are looking for a few unusual events buried in a large dataset consisting of almost all usual events.

The atmospheric process is represented by a time series collected under the auspices of the Atmospheric Radiation Measurement (ARM) project [11]. The time series are measurements of liquid water content of the atmosphere near Oak Ridge, Tennessee, over a period of 257 days beginning in March of 1994. The observations are taken at 20 second intervals but contain gaps, some over a day in duration. A segment of this time series is also used in Section 3.4. The segment considered here contains 122,786 observations covering a little over 28 days. It is the longest segment without any major gaps. The background process in this case corresponds to a relatively clear day with dry conditions. Perturbations include cloud, rain, and fog events as well as some instrument malfunction events. An important feature of the ARM time series is that the perturbations occur on several scales.

The second time series is a segment taken from one channel of a sixteen channel electroencephalogram (EEG) record for an epileptic patient [18]. The complete EEG record is 23 minutes, recorded at the sampling rate 512 *Hz*, and it includes a seizure. A 90 second segment, which occurs well in advance of the recorded seizure, is used for the analyses presented here. EEG records typically include a great deal of “artifact,” representing head movement, eye movement, muscle tension, grinding teeth, etc., in addition to unmasked neurophysiological activity. If we associate neurophysiological activity with the background process, then artifact is a perturbation relative to that background process.

In addition to the difference in fundamental form of the four techniques, there is also a difference in the windowing used to do the analyses presented here. The results presented for the regression model in Section 6.1 use non-overlapping windows of size $L = 50$, which corresponds to the segmentation Eq. 1. The analysis produces a result that applies to each individual window of data. The stochastic process model described in in Section 6.2 utilizes a moving window with structure, and the analysis results use a window of size $L = 405$. The segmentation is in keeping with that described by Eq. 2. In this case, the analysis produces a statement regarding the middle portion of the window of length 5, which is buffered on either side by pieces of length 200. The dynamic linear model discussed in Section 6.3 does not use a segmentation that fits either segmentation description of Eqs. 1 and 2. Rather, the window grows one observation per step, and a statement is made at an observation using all past history and a single future observed value. The fourth method, a nonlinear dynamical process model in Section 6.4, uses segmentation directly in keeping with the moving window paradigm of Eq. 2. However, a conclusion is reached about a collection of windows, which can vary in number, rather than about individual segments. The window length used here for both time series is approximately $L = 20$.

Color figures (Fig. 9 through 22) of this section appear in Appendix A.

6.1. Local Regression Models

We begin with the ARM data segment. It is partitioned into intervals of 50 observations giving a total of 2,455 intervals. As we discussed in Section 3 we fit a local regression model to each of the intervals. A quadratic model produces five values that characterize each interval: the three model coefficients and two measures of lack of fit (l_∞ and l_2 norms of residuals). This gives a data set of 2,455 observations on the five variables.

We can define a perturbation interval in terms of these five variables. Certainly intervals with a poor quadratic fit (high l_∞ and l_2 norm of residuals) can be considered to contain perturbations, but also intervals with unusually steep slope or a strong quadratic coefficient can be considered as parts of perturbations. In the ARM data segment we take the extreme (large in absolute value) 15% in any of the four variables as perturbation intervals. A plot of the time series is in Fig 9 showing the intervals without perturbations in black and intervals with perturbations in color. A more compute intensive but probably better separation could be obtained with multivariate density estimation techniques (see [26]). The 15% cutoff quantile was obtained by trial and error, but it can also be estimated from data. For example, given a sufficiently smooth unimodal density estimate for the l_2 variable, $\hat{f}(x)$, the quantile x_q that determines the extremes as perturbations can be estimated as

$$x_q = \arg \max_{\{x|\hat{f}'(x)<0, \hat{f}''(x)>0\}} \hat{f}''(x),$$

where $\hat{f}'$ and $\hat{f}''$ are the first and second derivatives of f. That is, we can take the point at which the smooth density estimate most rapidly "flattens out" into a tail.

We use a model based clustering method (see [4]) to cluster the intervals containing perturbations. This clustering method is well suited for overlapping ellipsoidal clusters of varying sizes and orientations. We use all five variables in clustering the intervals into four clusters, which are represented by red, magenta, green, and cyan in Fig. 9. Fig. 10 shows two scatterplots selected from the five clustering variables (l_2 vs. linear coefficient and l_2 vs. constant coefficient) that give a good view of the clusters. Most of the intervals without perturbations (black) are hidden behind the magenta cluster. Some observations from the two figures are:

- The red cluster represents intervals that contain a single transient.
- The magenta cluster represents perturbations on a small scale.
- The green and cyan clusters represent perturbations on a larger scale.

The EEG data segment is partitioned into intervals of 100 observations giving a total of 460 intervals. We fit a local quadratic model to each interval and keep the five values (three model coefficients and two measures of fit) that characterize the interval. This give a data set of 460 observations on five variables.

Similarly to the ARM data, we take the extreme 20% in any of four of the variables as perturbation intervals. A plot of the time series is in Fig 11 showing the intervals without perturbations in black and intervals with perturbations in color. Fig. 12 shows two scatterplots selected from the five clustering variables (l_2 vs. linear coefficient and l_2 vs. quadratic coefficient) that give a good view of the clusters. Some observations from the two figures are:

- magenta cluster represents mostly intervals with high variability;
- cyan cluster represents mostly intervals with medium variability;
- green cluster represents mostly intervals with convex curvature and good quadratic fit;
- red cluster represents mostly intervals with concave curvature and good quadratic fit.

6.2. A Stochastic Process Model

The method referred to here as "the changepoint detection method" is introduced and described in detail in Downing, et al.[10]. Here, we shall provide only a brief description. The method is based on a working assumption that Y is generated as the realization of a stationary Gaussian stochastic process. The method is intended to detect change points and outliers, where these are loosely defined to be any time at which the mean or covariance function changes, and one or a very few contiguous data points which would have very low probability under models which fit the surrounding data well, respectively.

Changing notation slightly from the above arguments, let Y_P denote the collection or vector of data values generated in a contiguous interval of x values, i.e. $P = [x_1, x_2], x_1 < x_2$. A "prediction" or "forecast" of Y_P, based on data which are located before P in time, can be constructed via the commonly used conditional approach to Gaussian time series. For specified parameter values, a commonly used forecast based on data from a conditioning interval C, with $C < P$, is the mean of the conditional Gaussian density

$$\phi(Y_P|Y_C),$$

sometimes called the predictive density in Bayesian applications. Rather than specifying parameter values arbitrarily, we take a more expedient route of estimating them from data in an estimation interval $E < P$, resulting in a predictive Gaussian density

$$\phi_E(Y_P|Y_C).$$

A measure of the quality of this prediction is its evaluation at the observed value of data in P.

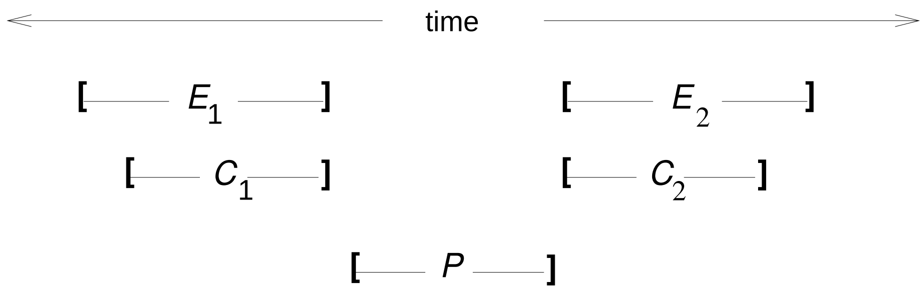

Figure 7: Intervals of the stochastic process model.

The basic idea behind our approach to detecting a change point is the comparison of two predictions Y_P, one a "forecast" as described above, and the other a "backcast" based on data from intervals which are subsequent to P in time. Formally, for a interval P of data, we also select intervals:

$$C_1 < P, E_1 < P$$

and

$$C_2 > P, E_2 > P.$$

The statistic is then defined as the log of the ratio of the larger to the smaller predictive densities, evaluated at the observed value of Y_P. Conservative critical values can be derived as described in Downing, et al.[10].

Fig. 7 displays graphically the relationships which exist between these intervals. Although in principle C_1 could appear anywhere before P in time, the predictor is most strongly influenced by the conditioning data if the two intervals are close together since, at least in many cases, correlations between data in the two intervals are strongest in this case. Also, it is reasonable that the estimation intervals should be relatively large, since parameter estimates should be reliable. Finally, because our approach assumes that the forecast and backcast should be equally reliable when changes or outliers do not occur, symmetry of size and placement of the estimation intervals and conditioning intervals around P is desirable. When the time series is "well behaved" without change points or outliers, values of the statistic should be small since both forecast and backcast would be expected to predict about equally well. On the other hand, if a change point exists in E_1 (E_2), i.e. the mean or covariance function changes somewhere in this interval, the forecast (backcast) should suffer due to loss of accuracy in the process parameters. Similarly, if an outlier exists in C_1 (C_2), the forecast (backcast) should also lose accuracy. (Both predictions will be relatively poor if the change point or outlier is actually in P, but this would not necessarily affect the comparative form of the statistic.) Hence a large value should be interpreted as evidence for a change or outlier in at least one of the five data intervals.

The procedure described above is used by "sliding" the arrangement of 5 intervals through the entire time series, evaluating the statistic at each time-step increment. Intervals of time over which the statistic is relatively large are flagged as containing potential change points or outliers, and subsequently reviewed graphically by the investigator to characterize their meaning in the context of the problem.

ARM Data Analysis

In this analysis, shown in Fig. 13, the intervals labeled by the examination of first differences (blue on the graph) appear to be associated with more dramatic changes in behavior than those labeled when the untransformed data are analyzed (yellow). In some cases, the untransformed analysis marks very early and late phases of apparent perturbations, while the differenced analysis marks the central interval (e.g. at approximately 64.4 and 73.8 on the time scale). This may reflect greater power to detect changes of level than of slope, but this is difficult to verify at the scale of the graph, particularly in the presence of obviously changing variability. Additionally, there are a number of intervals flagged by both analyses (green) where changes are not visually apparent at the plotted scale.

EEG Data Analysis

The plotted results of the Bayesian changepoint method in Fig. 14 are difficult to interpret graphically in this application. The analysis of untransformed data (yellow) seems more sensitive than that of first differences (blue) in this case. The marked intervals are much shorter here than in the ARM example, and a number of visually apparent changes are either not marked or are flagged over such short intervals that they are not visible on the plots. Results might have more closely reflected the visual impression given by the graph if longer window widths had been used.

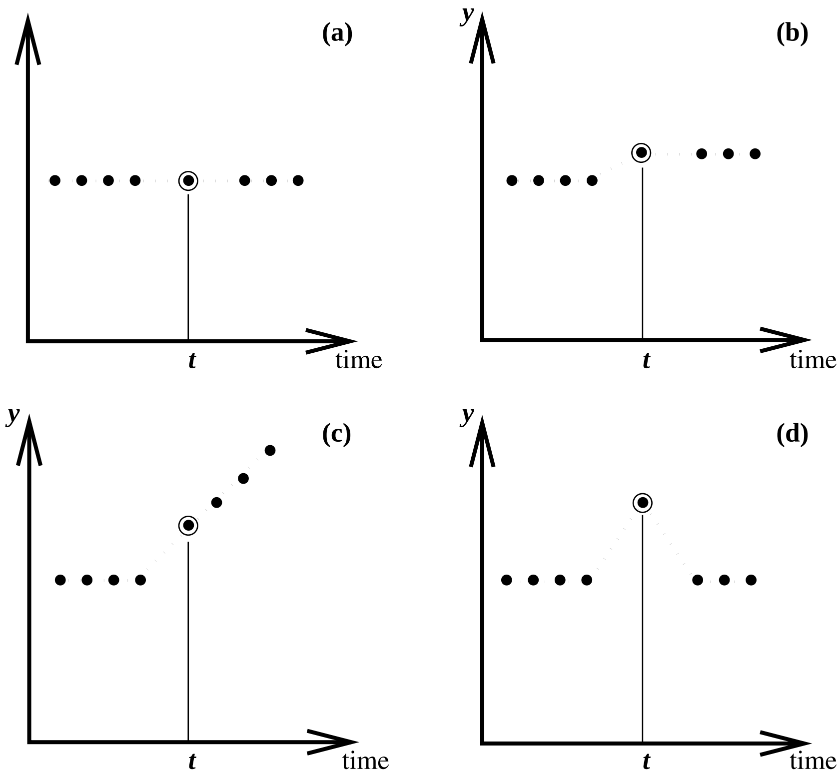

Figure 8: Four states of the Dynamic Linear Model: (a) no change, (b) level change, (c) slope change, and (d) transient change.

6.3. Dynamic Linear Model

Gordan and Smith [14] introduce a dynamic linear model [31] for detecting changes in "state" to time series data. Four distinct conditions in state are accounted for in the implementation of that model used for this study:

$j = 1$: no change in state
$j = 2$: level change in state
$j = 3$: slope change in state
$j = 4$: transient change in state

These are depicted in Fig. 8.

Letting $\{y_i\}_{i=1}^{\infty}$ denote the time series, for each j-value the model is defined by

$$y_i \;=\; \mu_i + \delta y_i^{(j)},$$

$$\begin{aligned} \mu_i &= \mu_{i-1} + \beta_i + \delta\mu_i^{(j)}, \\ \beta_i &= \beta_{i-1} + \delta\beta_i^{(j)}, \end{aligned}$$

where μ_i is level, β_i is level change due to slope or linear trend, and $\delta y_i^{(j)}$, $\delta\mu_i^{(j)}$, $\delta\beta_i^{(j)}$ are random variables such that

$$\begin{aligned} \delta y_i^{(j)} &\sim N(0, \lambda^{-1} R_y^{(j)}), \\ \delta\mu_i^{(j)} &\sim N(0, \lambda^{-1} R_\mu^{(j)}), \\ \delta\beta_i^{(j)} &\sim N(0, \lambda^{-1} R_\beta^{(j)}). \end{aligned}$$

Note that the variances in the normal distributions used to model fluctuations are state dependent. The state dependencies are qualitatively described as follows:

j	$R_y^{(j)}$	$R_\mu^{(j)}$	$R_\beta^{(j)}$
1	≈ 1	≈ 0	≈ 0
2	≈ 1	$\gg 1$	≈ 0
3	≈ 1	≈ 0	$\gg 1$
4	$\gg 1$	≈ 0	≈ 0

For each value of j, the model equations can be expressed in compact form as

$$\begin{aligned} y_i &= H_i\theta_i + \delta y_i^{(j)}, \\ \theta_i &= G_i\theta_{i-1} + \delta\theta_I^{(j)} \end{aligned}$$

where

$$\delta y_i^{(j)} \sim N(0, \lambda^{-1} R_y^{(j)}), \quad \delta\theta_i^{(j)} \sim N(0, \lambda^{-1} R_\theta^{(j)}),$$

and

$$\begin{aligned} \theta_i &= \begin{pmatrix} \mu_i \\ \beta_i \end{pmatrix}, \quad & H_i &= \begin{bmatrix} 1 & 0 \end{bmatrix}, \\ G_i &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad & R_\theta^{(j)} &= \begin{bmatrix} R_\mu^{(j)} + R_\beta^{(j)} & R_\beta^{(j)} \\ R_\beta^{(j)} & R_\beta^{(j)} \end{bmatrix}. \end{aligned}$$

Further, let $M_i^{(j)}$, $j = 1, 2, 3, 4$, denote state at time i and let $D_i = \{y_k\}_{k=1}^{i}$. Gordan and Smith assume that

$$p(\theta_0) \sim N(m_0, \lambda^{-1}C_0)$$

for some given m_0 and C_0. Further, they assume posteriors

$$\begin{aligned} p(\theta_{i-1}|D_{i-1}, \lambda, M_{i-1}^{(j)}) &\sim N(m_{i-1}^{(j)}, \lambda^{-1}C_{i-1}^{(j)}), \\ p(\lambda|D_{i-1}, M_{i-1}^{(j)}) &\sim G(\frac{1}{2}n_{i-1}, \frac{1}{2}r_{i-1}^{(j)}), \end{aligned} \tag{52}$$

where $G(\cdot,\cdot)$ is the Gamma distribution and initially $p(\lambda) \sim G(1/2n_0, 1/2r_0)$. Then, given the new observation y_i, Bayes's Theorem with the model leads to prescriptions for the posteriors $p(\theta_i|D_i, \lambda, M_i^{(j)})$ and $p(\lambda|D_i, M_i^{(j)})$. At this point, Gordan and Smith calculate

$$O_i^{(j)} = p(M_{i-1}^{(j)}|D_i),$$

that is, the probability that at time $i-1$ the state is $M_{i-1}^{(j)}$ given the data through time i. Gordan and Smith refer to these as *one-step-back* probabilities. The results presented below are obtained as follows: If $O_i^{(1)} < .05$, then there is a change in state at time $i-1$. The type of change in state is then specified to correspond to the largest value $O_i^{(j)}$, $j = 2, 3, 4$.

Implicit in the expressions Eq. (52) is knowledge of the observation variance, $V = \lambda^{-1}R_y$, and the system evolution covariances, $W_{i-1} = \lambda^{-1}C_{i-1}^{(j)}$. In practice, both of these quantities are unknown and difficult to specify. The Bayesian paradigm can be applied to updating the variances, but the system evolution covariance matrix updating can become very difficult from an analytical standpoint. Fortunately, this problem can be circumvented by a concept called *information discounting.* In general, as information ages it becomes potentially less useful and in some sense its value diminishes. The aging process is modeled in the dynamic linear model through the system evolution. This can be seen by comparing the prior variance of the system state at time i to that associated with the posterior variance of state. We see that

$$V(\theta_i|D_{i-1}) = V(\theta_{i-1}|D_{i-1}) + W_i$$

and that the increase in variance at time i represents our loss of information due to the advancement of time. If we view W_i as a proportional increase in variance, we can set

$W_i = \delta\, V(\theta_{i-1}|D_{i-1})$, which leads to

$$V(\theta_i|D_i) = (1+\delta)\, V(\theta_i|D_{i-1}).$$

Normally, information is measured in terms of precision, the inverse of variance. Working in terms of information, we can write

$$\begin{aligned} I(\theta_i|D_{i-1}) &= V^{-1}(\theta_i|D_{i-1}) \\ &= (1+\delta)^{-1}\, V^{-1}(\theta_{i-1}|D_{i-1}) \\ &= (1+\delta)^{-1}\, I(\theta_{i-1}|D_{i-1}). \end{aligned}$$

On this information scale, the discount factor $\lambda = (1+\delta)^{-1}$ varies over $[0,1]$. For a discount factor value $\lambda \in (0,1]$, the information loss through the evolution process is summarized as

$$V(\theta_i|D_{i-1}) = \lambda^{-1}\, V(\theta_{i-1}|D_{i-1}).$$

If we model the discount factor as a Gamma random variable, then the updating equations for the variance terms are quite tractable since the Gamma distribution is the natural conjugate prior for the Normal distribution.

ARM Data Analysis

The result of applying the linear dynamic model to detect changes in level and slope as well as transient effects was very successful. Fig. 15 uses the following color code:

level change	=	red
slope change	=	green
transient change	=	blue

There are a large number of points where the measurement value precipitously drops to zero. All of those points are detected and they are almost always identified as transients, which is correct. There are a few cases, however, where such points are classified as level changes. In several segments there are clear level changes, where the measurement either drops or rises and then gradually returns to the measurement value preceding the change. Slope change is detected least often, and where it does occur it appears to be associated with segments that have an erratic, rapidly varying appearance. In fact, because of that erratic structure, those segments having slope change points typically have a mixture of all three types of change. Consider the observation period $84.9 - 85.3\,Days$. For the scale used to display the data in Fig. 15, there do not appear to be any change points in that segment. However, the technique does mark as change points several pieces in that segment, which on a magnified scale are observed to be correctly identified.

EEG Data Analysis

Fig. 16 displays the results of the EEG data analysis. The same color code is used in that figure as is used to display the ARM results. The EEG time series is much more variable than the ARM data and some significant excursions occur. Most of those excursions are coded green, implying a slope change, and the figure clearly displays that a change in direction takes place. It is an interesting feature of the EEG data that these trends are rapidly changing with time, evidenced by clusters of green dots in the figure. Nested within some of those clusters of green are also points coded blue, indicating a level jump in the data. The only transient events occur in the time span $348.0 - 348.5\,s$, where extreme activity takes place. The segment $347.5 - 350.0\,s$ is atypical and is liberally sprinkled with green and red dots, indicating a rapid mixture of change.

6.4. A Nonlinear Dynamical Model

We consider a method for identifying significant transients in time series data based on the theory of nonlinear dynamical processes. We are observing a process that resides in a relatively quiescent mode for significant periods of time, but aperiodically becomes excited for finite periods before relaxing to the background, relatively quiescent state.

We want to give a qualitative description of the technique applied here to address the detection problem. Suppose the process we are observing is represented by a trajectory in state space. Further, suppose most of the time that trajectory is confined to a limited region in state space, which will be referred to here as an attractor. Now, suppose occasionally the trajectory leaves that attractor and moves about in an extended region of state space before falling back to the attractor. Such a trajectory segment will be called a "perturbation." We shall say the attractor corresponds to the "background" process represented by the time series. If that attractor is moving or changing with time, we shall say that background process is non-stationary.

We refer to Section 5.4 for the definition of notation used here. Suppose we have a "training" set of data, $\{y_i\}_{i=1}^N$, representative of the background process, and suppose we have gone through the procedure of constructing the representation $(\tilde{A}_\theta, \tilde{f}_\theta, \tilde{p}_\theta)$ of the process corresponding to that training set. Recall that in constructing that representation we select values for the parameter pair (n, k), then define the trajectory

$$\{\mathbf{y}_i\}_{i=1}^N \subset E_y^n,$$

where

$$\mathbf{y}_i = (y_i, y_{i+k}, y_{i+2k}, \ldots, y_{i+(n-1)k})^T,$$

and then carry out the transformation Eq. (48)

$$\Psi : E_y^n \to E_\theta^n,$$

thus producing the trajectory

$$\{\Psi(\mathbf{y}_i)\}_{i=1}^N = \{\theta_i\}_{i=1}^N \subset E_\theta^n.$$

Further, recall that a good choice for the parameter pair (n, k) produces eigenvalues $\{\lambda_j\}_{j=1}^n$ Eq. (47) that decrease rapidly, so that in general the model Eq. (49)

$$\mathbf{y}_i = \sum_{\alpha=1}^{n} \theta_{i\alpha} \psi_\alpha$$

can be replaced by the model Eq. (50)

$$\mathbf{y}_i = \sum_{\alpha=1}^{m} \theta_{i\alpha} \psi_\alpha + \epsilon_i,$$

where $m < n$ is the range of significant eigenvalues and ϵ_i is a random noise term used to model the insignificant eigenvalue components.

The problem at hand is to identify perturbations relative to some background process. In that circumstance, we are not trying to accurately model the background process but, rather, are trying only to separate the unusual from the usual. To accomplish that end, we propose to do what is equivalent to using a weaker regression model than that described above Eq. (50), where $m < n$ represents the range of significant eigenvalues. Define the projection $P_\theta^{m_1,m_2} : E_\theta^n \to E_\theta^{m_2-m_1+1}$ by

$$P_\theta^{m_1,m_2}(\theta) = (\theta_{m_1}, \ldots, \theta_{m_2})^T \in L\{\psi_j\}_{j=m_1}^{m_2},$$

where $L\{\psi_j\}_{j=m_1}^{m_2}$ is the linear subspace of E_θ^n spanned by the eigenvectors $\{\psi_j\}_{j=m_1}^{m_2}$. Further, define

$$\tilde{B} = P_\theta^{m_1,m_2}(\{\theta_i\}_{i=1}^N).$$

Thus, $\tilde{B}$ is the projection into the $(m_2 - m_1 + 1)$-dimensional subspace $L\{\psi_j\}_{j=m_1}^{m_2}$ of the trajectory $\{\theta_i\}_{i=1}^N$ in E_θ^n constructed from the training set $\{y_i\}_{i=1}^N$. We assume that the background process projects to a relatively small, dense region $\tilde{B}$ for a small value of $m_2 - m_1 + 1$.

The assumptions described above concerning the background process versus perturbations implies that the region $\tilde{B}$ is a concentrated region in state space associated with usual data segments and that unusual segments of the time series will produce trajectory segments in $L\{\psi_j\}_{j=m_1}^{m_2}$ that move outside $\tilde{B}$. Let T_s be a characteristic time scale associated with the background process. We define a background event, or usual segment of the observed time series by conditions on the reconstructed trajectory segment in E_θ^n. Define

$$\Gamma_j = \{\theta_i\}_{i=i_j}^{i_j+l_j-1},$$

so that l_j is the length in time steps of Γ_j. If

$$\begin{aligned} P_\theta^{m_1,m_2}(\theta_{i_j-1}) &\notin \tilde{B}, \\ P_\theta^{m_1,m_2}(\Gamma_j) &\subset \tilde{B}, \\ P_\theta^{m_1,m_2}(\theta_{i_j+l_j}) &\notin \tilde{B}, \\ l_j \times t_s &\geq T_s, \end{aligned}$$

where t_s is the sample time for the time series, then the trajectory segment Γ_j is a usual event and the time series segment corresponding to Γ_j is a usual segment. In turn, we define the trajectory segment

$$\Delta_j = \{\theta_i\}_{i=i_j+l_j}^{i_{j+1}-1}$$

separating Γ_j and Γ_{j+1} to be the j-th perturbed segment. The length in time steps of Δ_j is

$$p_j = (i_{j+1} - 1) - (i_j + l_j) + 1 = i_{j+1} - (i_j + l_j).$$

Note that this definition allows a perturbed trajectory segment to pass through the region $\tilde{B}$ so long as the time it takes is less than the time scale T_s. The time series segment corresponding to Δ_j is, by definition, a perturbation, or unusual segment.

Referring to Downing, et al.[9] for details, we find that the atmospheric(ARM) and neurophysiological(EEG) data sets can be modeled using the parameter pair values

$$\begin{aligned} \mathrm{ARM} : (n, k) &= (8, 3), \\ \mathrm{EEG} : (n, k) &= (9, 3). \end{aligned}$$

Further, for both example time series we find that the dominate structure of the background process is included in the subspace of E_θ^n spanned by the first three eigenvectors, $E_\theta^3 = L\{\psi_j\}_{j=1}^3$, so that $m = 3$ in the regression model Eq. (50). If there is a long term trend in the background process, we expect that to be reflected primarily in the first coefficient, θ_1, of the regression model. Consequently, we use the coefficients (θ_2, θ_3), which is equivalent to setting

$$(m_1, m_2) = (2, 3).$$

Thus, $\tilde{B}$ is the projection into the two dimensional subspace $L\{\psi_j\}_{j=2}^3$ of the trajectory $\{\theta_i\}_{i=1}^N$ in E_θ^n constructed from the training set $\{y_i\}_{i=1}^N$. Finally, we find that appropriate time scales for defining usual events are

$$\begin{aligned} \mathrm{ARM} : T_s &= 180 \times t_s = 1hr, \\ \mathrm{EEG} : T_s &= 100 \times t_s \approx .2s. \end{aligned}$$

For the examples presented here clustering of perturbations, that is, the division of unusual events into distinct classes, is done using only the duration of an event. From the definition of the j-th perturbed trajectory segment, Δ_j, the time duration of Δ_j is

$T_j = p_j \times t_s$. Three classes of perturbations are distinguished for each of the two examples.

	EEG	ARM
1	$.0s < T_j < .1s$	$0hr < T_j < 2hr$
2	$.1s \leq T_j < .2s$	$2hr \leq T_j < 7hr$
3	$.2s \leq T_j$	$7hr \leq T_j$

The time scales used to define classes, $.1s$ for EEG data and $1hr$ for ARM data, are meaningful, but the decision to display just three classes is arbitrary.

The analysis results for the ARM example are displayed in Figs. 17 and 18, and those for the EEG data in Figs. 19 and 20. Recall from the introduction to this section that the data selected for examples was chosen on the basis that they are rich in features. No effort was made, in particular, to insure the perturbation durations would be consistent with some preset distribution, say, a uniform distribution.

The ARM data has features that vary over a broad range of scales. Thus, the results illustrated in Fig. 17, which shows perturbations coded on the given time series, reveal segments that are clearly unusual to the eye, but also mark other segments that do not appear to be unusual. Note the corresponding scatter plot, Fig. 18. Recall that (θ_2, θ_3) is displayed in that plot. Qualitatively speaking, it appears the distributions of perturbed (θ_2, θ_3) values are approximately the same for the three classes, or temporal bands, described above. Scale differences are largely confined to θ_1, which if plotted appears virtually identical to the given time series Fig. 17.

Features in the EEG data are not as widely distributed across scales as the ARM data. Fig. 19 is the perturbation coded time series and Fig. 20 is the corresponding scatter plot. We find that the technique is very efficient at identifying artifact. Referring to Fig. 19, for example eye movement at times $328.5s, 338.0s$, and $343.0s$ is captured and muscle tension with mixed head movement is captured in the time periods $(311.5s, 314.5s)$ and $(348.0s, 350.0s)$. From the scatter plot Fig. 20, it appears the third class of perturbations dominates in the particular time series data used for this study. While the results for the first two classes are sparse, the distributions of perturbed values for (θ_2, θ_3) appear to be roughly the same for all three classes.

6.5. A Summary of Detection Analysis Results

The detection techniques described above fall into two groups, one generally referred to as *change point analysis* [22, 23, 6], and the second as *structural change analysis* [16]. The stochastic process model described in Section 6.2 and the dynamic linear model described in Section 6.3 are examples of change point analysis, and the local regression model presented in Section 6.1 and the dynamical systems model presented in Section 6.4 are examples of structural change analysis. The discussion here is limited to the ARM dataset results, but that is an arbitrary decision.

The local regression model and nonlinear dynamical systems model give very similar results for both example time series. While a little different in detail, both approaches identify a high density region in some parameter space and define usual events using that region. In the case of the local regression model, if the parameters of a given segment fall in the high density region, then that segment is, by definition, usual; else, that segment is unusual. If the segment is unusual, it is assigned to a class by some measure of the magnitude of the representation in parameter space. In the case of the dynamical systems model, a time scale is used in addition to the high density region, and a trajectory segment in parameter space, or state space, is defined to be usual if every point on the trajectory segment is in the high density region and the duration of the segment is at least as long as the specified time scale. A perturbed segment is simply one joining two usual segments, so that it is modeled as a trajectory in state space that leaves the high density region and moves about in state space before collapsing back to the high density region. A perturbed trajectory segment may at different times be far away from the high density region, near, but still outside that region, or even inside it. An unusual segment is assigned to a class using duration as a measure.

Consider the ARM data analysis results for local regression and dynamical systems. Referring to the coded time series Figures 9 and 17, the segment covering 65.0-66.0 *Days* provides a typical comparison. That segment of the time series includes parts that are classified in all four categories according to both methods, and in each case most of the 1 *Day* segment is unusual. For the dynamical systems approach, the initial .7 *Days* belongs to one class, the one of greatest duration, whereas that segment is distributed among all four classes used with the local regression model. The difference in classification is explained by the fact that for the dynamical systems approach the perturbed trajectory segment moves about in state space at various distances from the high density region at different times. If distance from the high density region were banded and coded in a way consistent with how the magnitude of a perturbed segment is coded using local regression, the figures would look virtually identical.

The large feature from the ARM data corresponding to the 65.0-66.0 *Days* segment is clearly observable by eye, but consider for example the time segment 78.4-79.0 *Days*. A magnified view of the coded time series for the two methods is given in Figs. 21 and 22. Again, the results are very consistent, taking into consideration the differences in defining usual and unusual described above.

The stochastic process model is designed to detect a change in the observed stochastic process. At a given stage, a decision is made about a test window of observation length $\delta i = 5$ using 200 observations on either side of that window to estimate parameters and windows of length $\delta i = 20$ immediately preceding and following the test window to make a conditional forecast and a conditional backcast about the process in the test window. In addition to the test window, the most critical observations at a stage are the conditioning windows, in this case of length $\delta i = 20$, on either side of the test window. If the forecast and backcast agree, and if those agree to a specified tolerance with the estimation of the process based on the test window alone, then there is no change.

Consider the ARM data analysis results using the stochastic process model Fig. 13. There are 4320 observations per day, so that .1 days corresponds very closely to the total number of observations used to analyze a test window. As just described, the number of most critical observations is $L = 5 + 2 \times 20 = 45$. Consider the bump in the interval $65.2 - 65.3\,Days$. Note that the bump itself is colored black, an indication of no change, while the preceding and succeeding feet of the bump are color coded green, indicating change. This makes sense because, for example, if we consider the preceding segment, in that segment the backcast includes critical observations from within the bump and, so, produces an estimate for the test window that does not agree with either the forecast or test window estimates. In comparison, for both the local regression model Fig. 9 and the nonlinear dynamical model Fig. 17, this bump is identified as a perturbation. Compare this bump to the one in the interval $73.8 - 73.9\,Days$. The explanation for determining change in the preceding and succeeding feet of the second bump is the same as for the first. But why is the second bump identified as change? Note the asymmetry in the second bump relative to the first, and compare the critical length $L = 45$ to the length in number of observations of the bump. On the leading face of the bump, it appears likely that the forecast, which includes information from the leading foot, and backcast, which includes information from the trailing face, disagree and that both disagree with the test window estimate. Again in comparison with the local regression model result Fig. 9 and the nonlinear dynamical model result Fig. 17, both of those models identify the second example bump as a perturbation. As a final example, consider the bump between $80.7 - 81.0\,Days$. The relatively flat section along the peak in the interval $80.7 - 80.8\,Days$ is almost all coded black, implying no change, and there is a segment on the trailing face at approximately $80.9\,Days$ that is also coded black. This third example bump is again identified as a perturbation by both the local regression model and the nonlinear dynamical model.

Consider next the dynamic linear model technique. We refer to the results for the ARM data Fig. 15 and look at the same three bumps that were used above to describe results for the stochastic process model. First, in general it appears that the dynamic linear model produces very localized results. For the first bump, the one in the interval $65.2 - 65.3\,Days$, changes in state are marked on the leading face near the peak and at the peak, but the trailing face together with preceding and following feet of the bump are each associated with no change in state. Compare this with the stochastic model analysis Fig. 13, which views this bump as a single process, but distinguished it from leading and following segments. The second example bump is in the interval $73.8 - 73.9\,Days$. Again, the leading face is identified with a change in state, but like the first example bump the trailing face together with the preceding and succeeding feet are determined to be in steady state. The stochastic process model associates the bump, including an interval containing it, with a change in process, although the process change for the feet is distinguished in type from that for the bump. The third example is the bump in the interval $80.7 - 81.0\,Days$. Note that a short piece of the leading edge is marked as a transition in state and the plateau, which is rugged, is also not in steady state. The leading and trailing faces, not counting the single point on the trailing face that is an instrumentation anomaly, are in steady state. This result is virtually opposite that of the stochastic process model, which in particular determines the plateau to be a single process.

7. Conclusions

Large data set analysis is a phrase that may appear at first to be simple to define, but on more careful examination it becomes somewhat elusive. Are we observing a process that really requires gigabytes, or terabytes of data to understand it? Or, did we design the experiment, or the data acquisition strategy, so carelessly that almost all the data is useless and our task is to weed out the little bit of what is useful? In the former case, we are forced to analyze the data as one piece, with no chance to first reduce the size of the problem through some sort of a data compression algorithm, and simply make a reasonably efficient use of available computing hardware by implementing classical analysis techniques in ways that take advantage of that hardware design. In the latter case, surely after a while we would discover the mistake and make a correction.

Providing a meaningful description of the phrase "large data set analysis" is, in fact, a problem, or a concept, that is currently receiving a lot of attention. In the present study we have implicitly selected one example of such a definition. Stated qualitatively, we suppose the natural state of the process being observed is not interesting and it is only rare, randomly occurring events, which we know very little about a priori, that we want to recognize and save. Clearly, one important use of algorithms implementing this sense of analysis would be to the design of data acquisition systems used for observing such processes.

Starting with the sense of "large data set analysis" given above, we have developed four distinct techniques for demonstration, using some fairly weak, and thus widely applicable, assumptions. In the following, we summarize the most important consequences of those assumptions and the most significant features of that collection of techniques. We refer to Section 6.5 for detailed information concerning performance characteristics of those methods.

1. "Standard" approaches such as regression analysis, Fourier analysis, Box-Jenkins' procedure, et al., which handle a data set as a whole, are not admissible for very large data sets for at least two reasons. First, even with computer hardware available today, including parallel processors and storage devices, there are no effective means for manipulating and analyzing gigabyte, or larger, data files. Second, in general it can not be assumed that a very large data set is "stable" by the usual measures, like homogeneity, stationarity, and ergodicity, that standard analysis techniques require.

Both reasons dictate the necessity to use "local" data analysis methods whereby the data is segmented and ordered, where order leads to a sense of "neighbor," and then analyzed segment by segment. The idea of local data analysis is central to the study reported here.

2. The expression "data segmentation" is, by itself, ill-defined, because there are many alternative strategies that the practitioner might imagine, depending on the application, for localizing the analysis and each one carries with it some sense of segmentation. The four local analysis methods described and applied in Section 6 determine two naive senses of segmentation, that of a partition into distinct, non-overlapping pieces, and that of a moving window. Two of the local analysis methods, those described in Sections 6.2 and 6.3, are recursive in form and intrinsically lead to moving window segmentations. The local regression method described in Section 6.1 and the dynamical systems method described in Section 6.4 are examples that can be modified to lead to either of those two senses of segmentations. The essential difference in local analysis results that can be expected from the two segmentations is described in Section 3.5.

3. Another central theme of the methods studied in this article is that we are separating the "unusual" from the "usual", so that there is a great deal of information, indeed, redundant information, about what is usual. This theme is also related to the issue of data segmentation, because we want the local analysis to be a low-order, computationally fast, and statistically robust technique for determining if a segment is usual. If the test for what is usual fails, the segment is, by definition, unusual. An important problem is optimizing the segmentation size relative to, first, the condition that what is usual is simple and, second, the condition that what is unusual is not lost in overwhelming usual information.

4. The methods described in this article are very universal and may be used with virtually no a priori information about the process represented by the data. Clearly, any independent information about the process that serves to distinguish between the usual and the unusual of interest, such as time scales for example, can, and should be used in a particular application. Indeed, steps should be taken to tailor the method to that application. This is why in this report we are describing our methods with two specific applications. The methods we have described have the property that they

can be tailored to reflect given information about the process under consideration and the objective of the analysis.

5. The collection of techniques is made up of two fundamentally different types, namely, structural change analysis and change point analysis. The structural change type identifies the whole of what is distinguished from the usual, while the change point type identifies where a transition from one process to another takes place. For example, if the usual is dry, clear atmospheric conditions and one kind of unusual is a steady rainfall, then the structural change analysis will separate the entire rainfall event, while the change point analysis will separate the transient from dry and clear to steady rainfall and the transient from steady rainfall to dry and clear. Each technique generates information that can be used to classify events.

6. The segmented and ordered data structure construct taken together with the local analysis philosophy lends itself directly to parallel computational implementation. The collection of techniques described in this study is extendible to multivariate form. One example where it appears reasonable to consider the use of parallel computing resources is for analyzing series of photographs for dynamic events. The analysis would involve mapping the high dimensional photographs to vectors of some relatively low dimension, followed by applying the multivariate form of one or more of the techniques described here.

8. References

[1] H.D.I. Abarbanel, R. Brown, J.J. SiDorowich, and L.S. Tsimring. The analysis of observed chaotic data in physical systems. *Rev. Mod. Phys.*, 65:1331–1392, 1993.

[2] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK Users' Guide, Second Edition.* SIAM, Philadelphia, PA, 1995.

[3] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley and Sons, New York, New York, 1958.

[4] Jeffrey D. Banfield and Adrian E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, 1993.

[5] J. Beran. *Statistics for Long-Memory Processes.* Chapman and Hall, New York, New York, 1994.

[6] B. E. Brodsky and B. S. Darkhovsky. *Nonparametric Methods in Change-Point Problems.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.

[7] D. S. Broomhead and G. P. King. Extracting qualitative dynamics from experimental data. *Physica*, 20D:271, 1986.

[8] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JRSS(B)*, 39:1–38, 1977.

[9] D. J. Downing, V. V. Fedorov, W. F. Lawkins, M. D. Morris, and G. Ostrouchov. Analysing perturbations and nonstationarity in time series using techniques based on the theory of chaotic nonlinear dynamical systems. Technical Report ORNL/TM-13115, Oak Ridge National Laboratory, Oak Ridge, TN 37831, 1995.

[10] D. J. Downing, W. F. Lawkins, M. D. Morris, and G. Ostrouchov. A method for detecting changes in long time series. Technical Report ORNL/TM-12879, Oak Ridge National Laboratory, Oak Ridge, TN 37831, 1995.

[11] D. O. E. Atmospheric radiation measurement program plan. Technical Report DOE/ER-0441, U. S. Department of Energy, Office of Health and Environmental Re-

search, Atmospheric and Climate Research Division, National Technical Information Service, 5285 Port Royal Road, Springfield, Virginia 22161, 1990.

[12] J.P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, 57:617, 1985.

[13] V. V. Fedorov, P. Hackl, and W. G. Muller. Estimation and experimental design for second kind regression models. *Informatik, Biometric und Epidemiologue in Medizin und Biologie*, 24:134–151, 1993.

[14] K. Gordan and A. F. M. Smith. Modeling and monitoring biomedical time series. *J. Amer. Stat. Assoc.*, 85:328–337, 1990.

[15] P. Grassberger, T. Schreiber, and C. Schaffrath. Nonlinear time sequence analysis. *Int. J. Bifurcation and Chaos*, 1:521–547, 1991.

[16] P. Hackl. *Statistical Analysis and Forecasting of Economical Structural Change.* Springer-Verlag, New York, 1989.

[17] D. M. Hawkins. *Identification of Outliers.* Chapman and Hall, New York, New York, 1980.

[18] L. M. Hively, N. E. Clapp, C. S. Daw, W. F. Lawkins, and M. L. Eisenstadt. Nonlinear analysis of eeg for epileptic seizures. Technical Report ORNL/TM-12961, Oak Ridge National Laboratory, Oak Ridge, TN 37831, 1995.

[19] R. P. Kanwal. *Linear Integral Equations, Theory and Technique.* Academic Press, New York, New York, 1971.

[20] A. N. Kolmogorov and V. A. Uspenskii. Algorithms and randomness. *Theory Probab. Appl.*, 32:389–412, 1988.

[21] W. F. Lawkins, C. S. Daw, D. J. Downing, and N. E. Clapp. The role of low-pass filtering in the process of attractor reconstruction from experimental chaotic time series. *Physical Review E*, 47:2520–2535, 1993.

[22] E.S. Page. Continuous inspection schemes. *Biometrika*, 1:100–115, 1954.

[23] E.S. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42:523–526, 1955.

[24] C. R. Rao. *Linear Statistical Inference and Its Applications.* John Wiley and Sons, New York, New York, 1973.

[25] Statistical Sciences. *S-PLUS Guide to statistical and mathematical analysis, Version 3.2.* StatSci, a division of MathSoft Inc., Seattle, 1993.

[26] David W. Scott. *Multivariate Density Estimation: theory, practice, and visualization.* John Wiley & Sons, Inc., New York, 1992.

[27] G. A. F. Seber. *Linear Regression Analysis.* John Wiley and Sons, New York, New York, 1977.

[28] A. Shen. Algorithmic complexity and randomness: Recent developments. *Theory Probab. Appl.*, 37:92–97, 1995.

[29] F. Takens. Detecting strange attractors in turbulence. In D. A. Rand and L.S. Young, editors, *Proceedings of the Warwick Symposium*, page 366, New York, 1981. Springer.

[30] H. Tong. *Non-Linear Time Series: A Dynamical Systems Approach.* Clarendon Press, Oxford, 1990.

[31] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models.* Springer-Verlag, New York, New York, 1989.

A. Color Figures

This appendix contains all color figures.

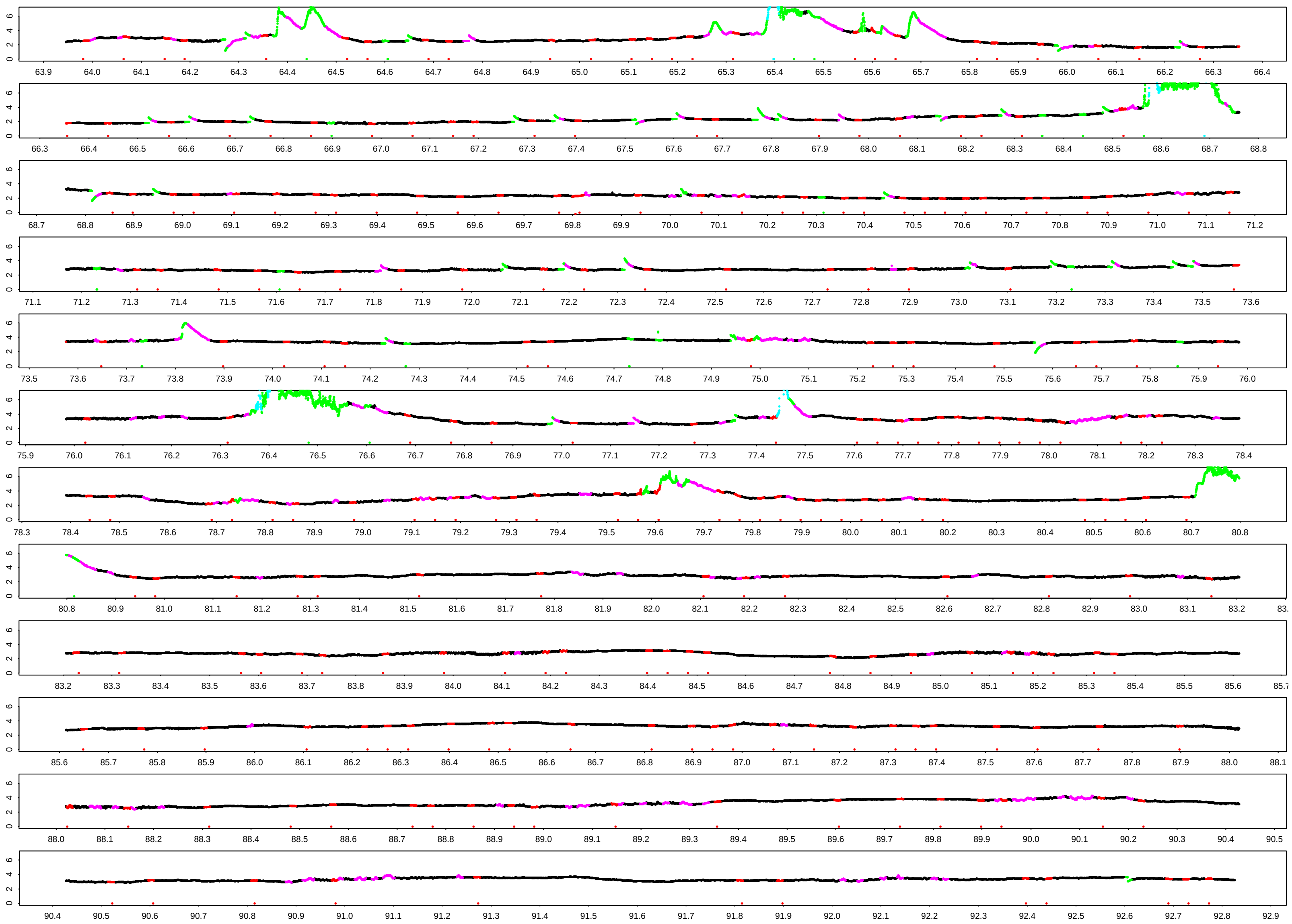

Figure 9: ARM time series segment (days vs. cm) with intervals colored according to clusters formed from parameters of local regression models. Perturbation free intervals are black. Other intervals are clustered into four groups denoted by red, magenta, green, and cyan.

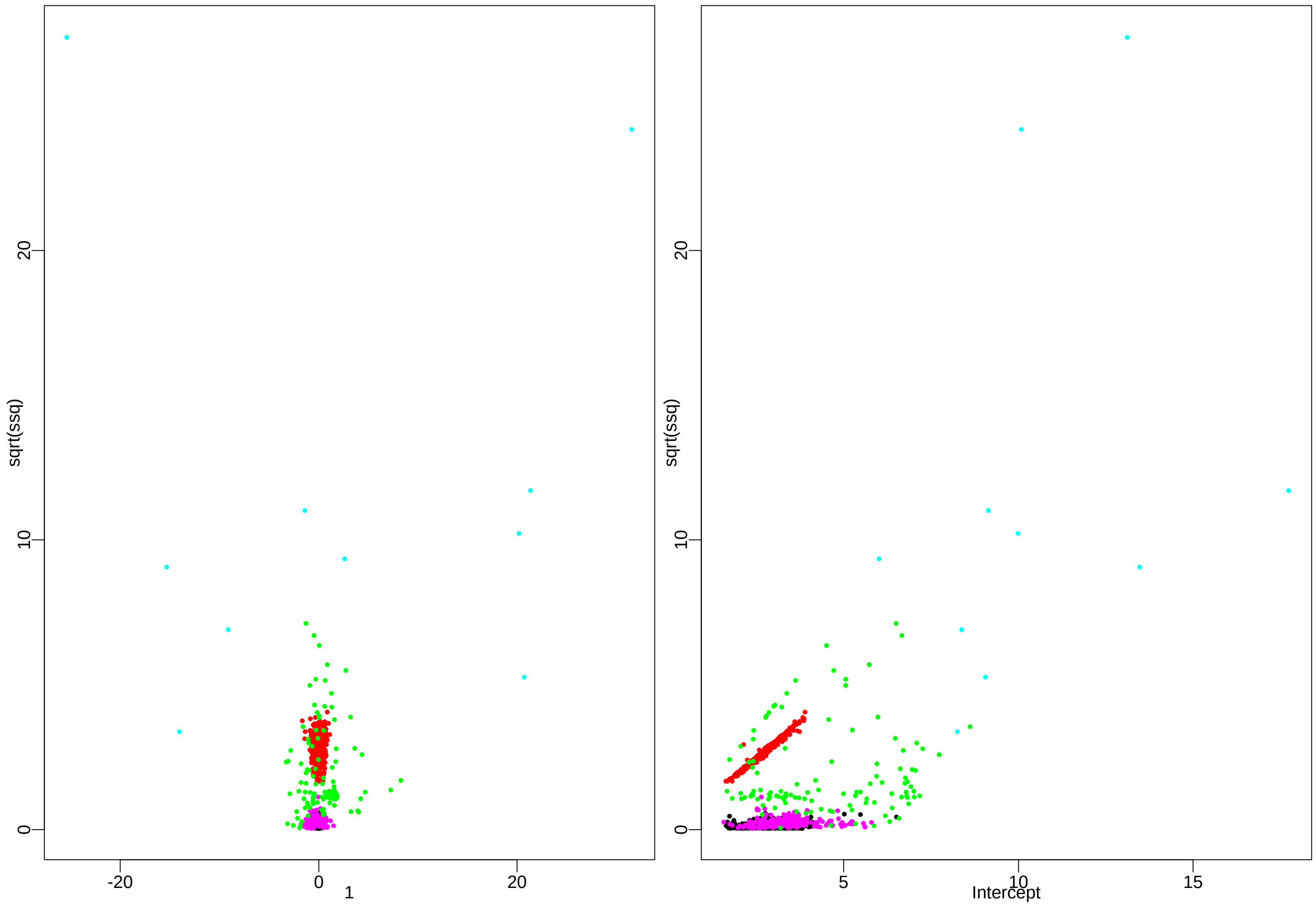

Figure 10: Clustered scatterplots of pairs of parameters from ARM time series intervals. Perturbation free intervals are black. Other intervals are clustered into four groups denoted by red, magenta, green, and cyan.

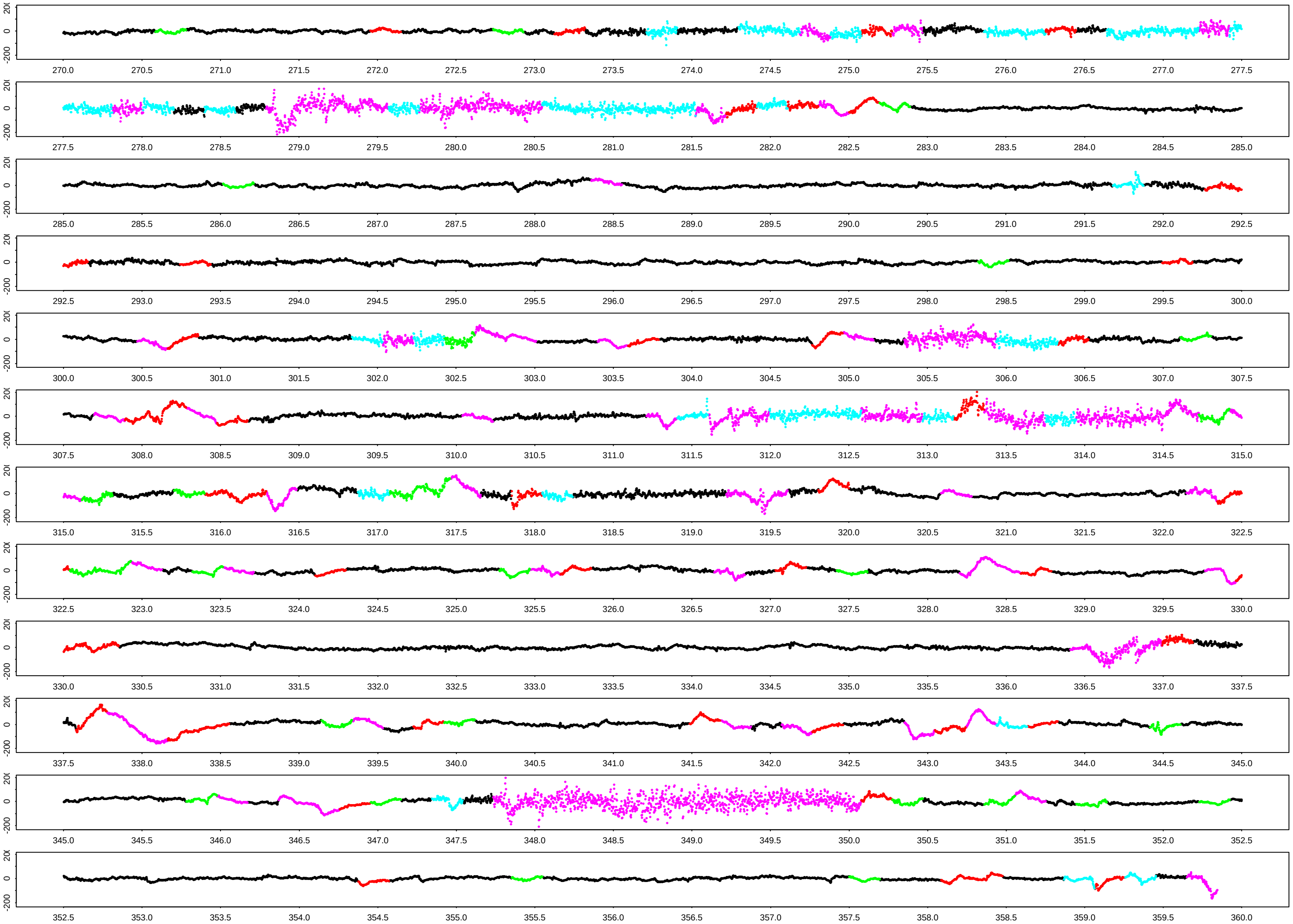

Figure 11: EEG time series segment with intervals colored according to clusters formed from parameters of local regression models. Perturbation free intervals are black. Other intervals are clustered into four groups denoted by red, magenta, green, and cyan.

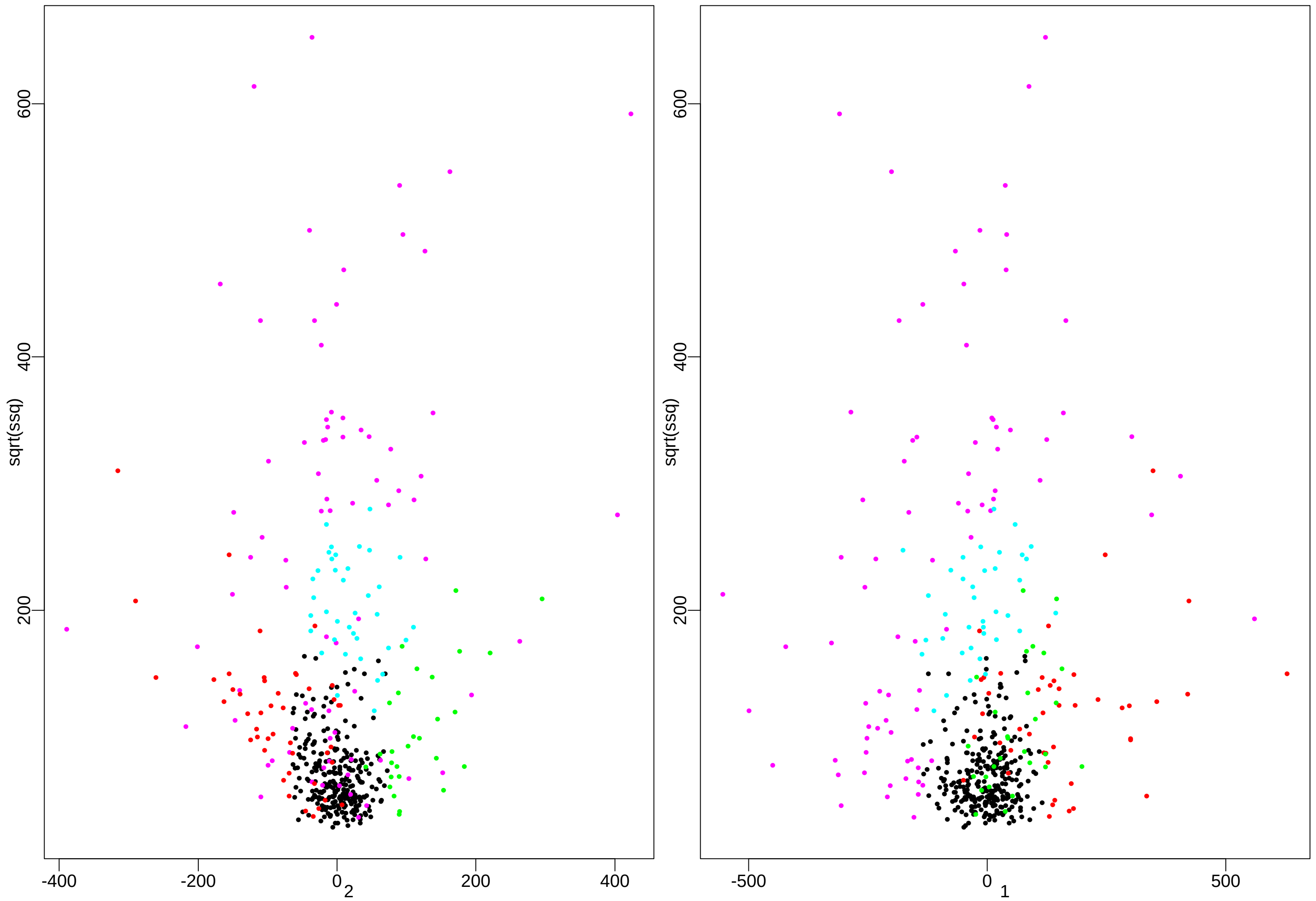

Figure 12: Clustered scatterplots of pairs of parameters from EEG time series intervals. Perturbation free intervals are black. Other intervals are clustered into four groups denoted by red, magenta, green, and cyan.

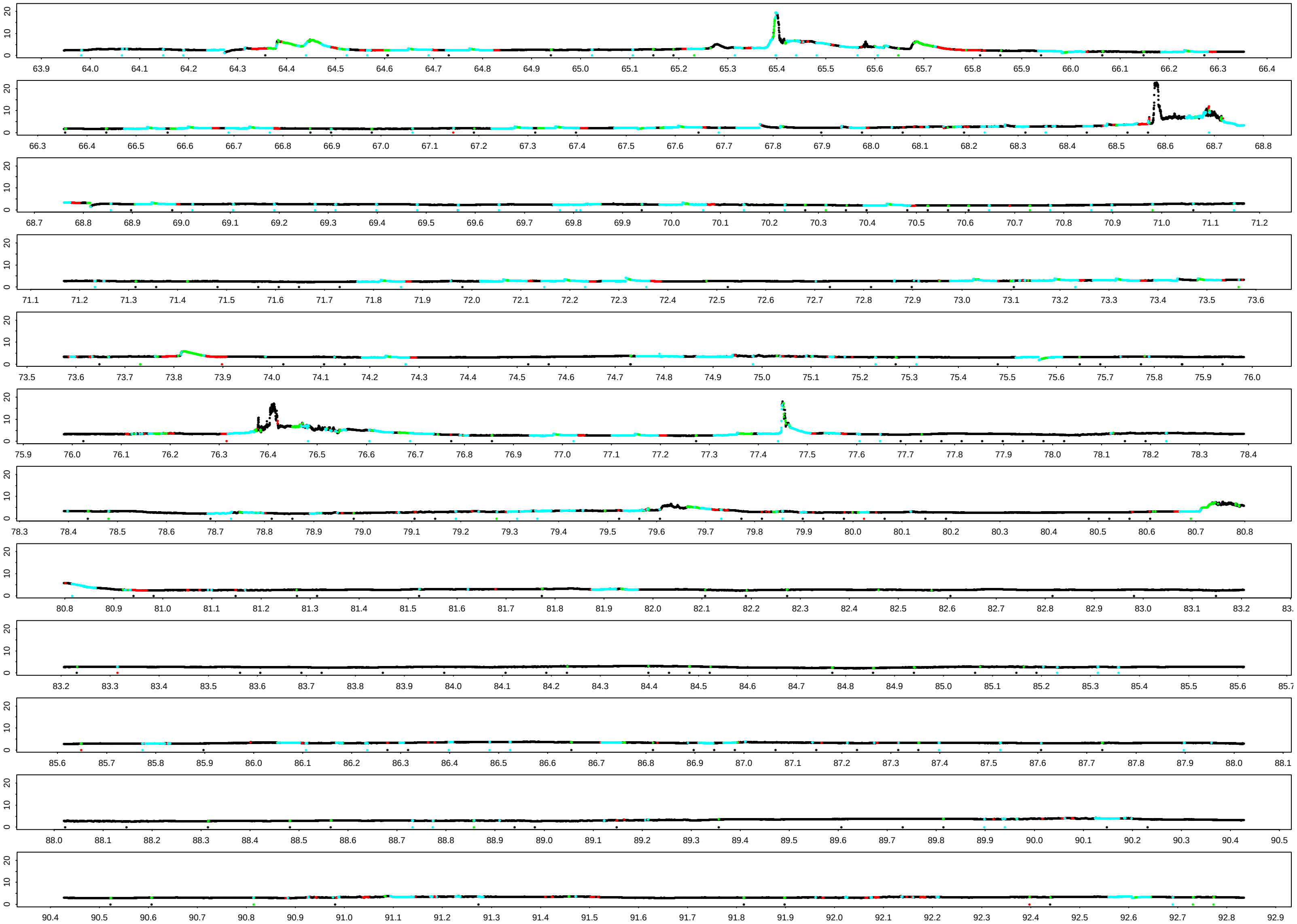

Figure 13: ARM time series segment (days vs. cm) with perturbations identified by a stochastic process model. Perturbation free portions are black.

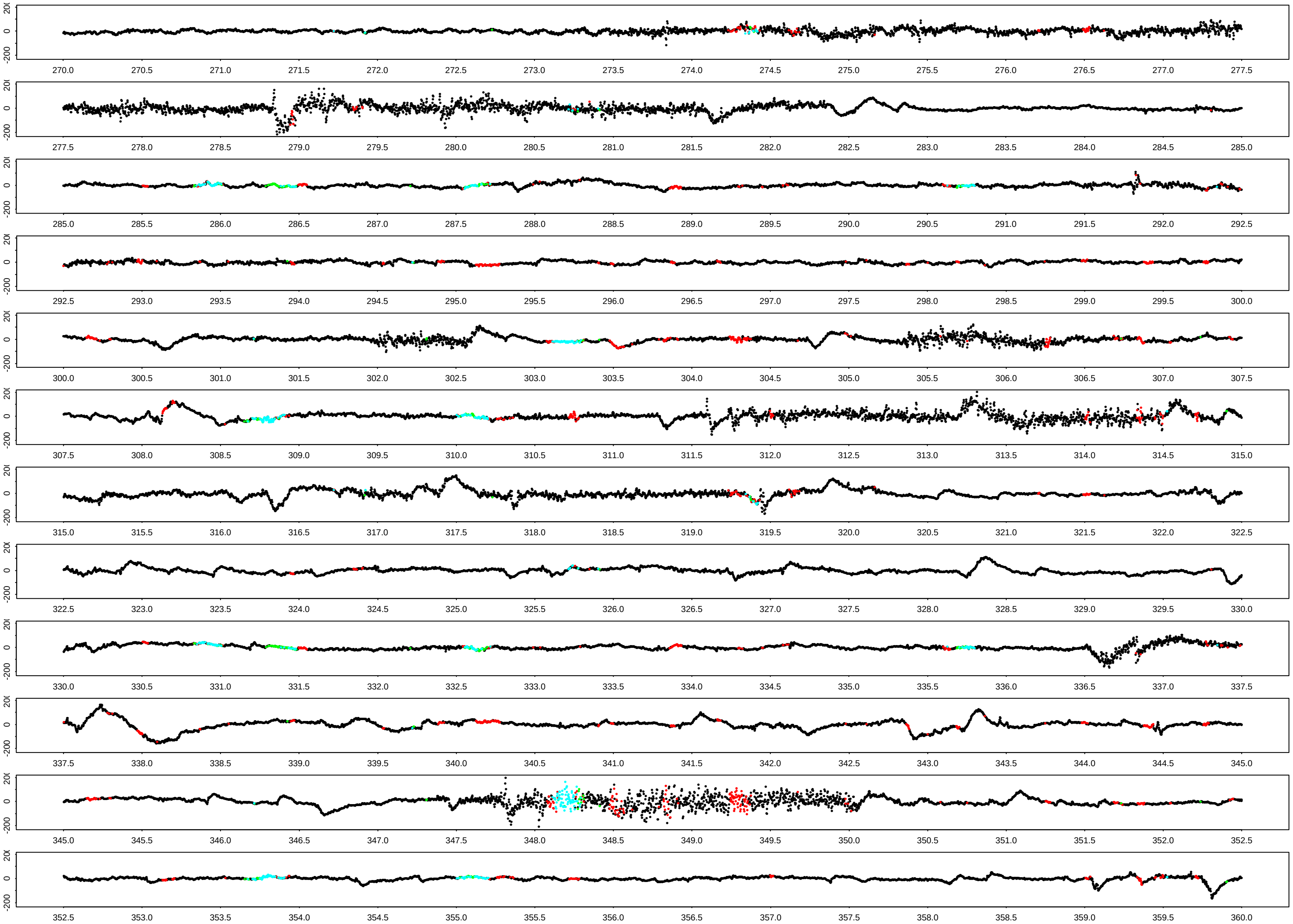

Figure 14: EEG time series segment with perturbations identified by a stochastic process model. Perturbation free portions are black.

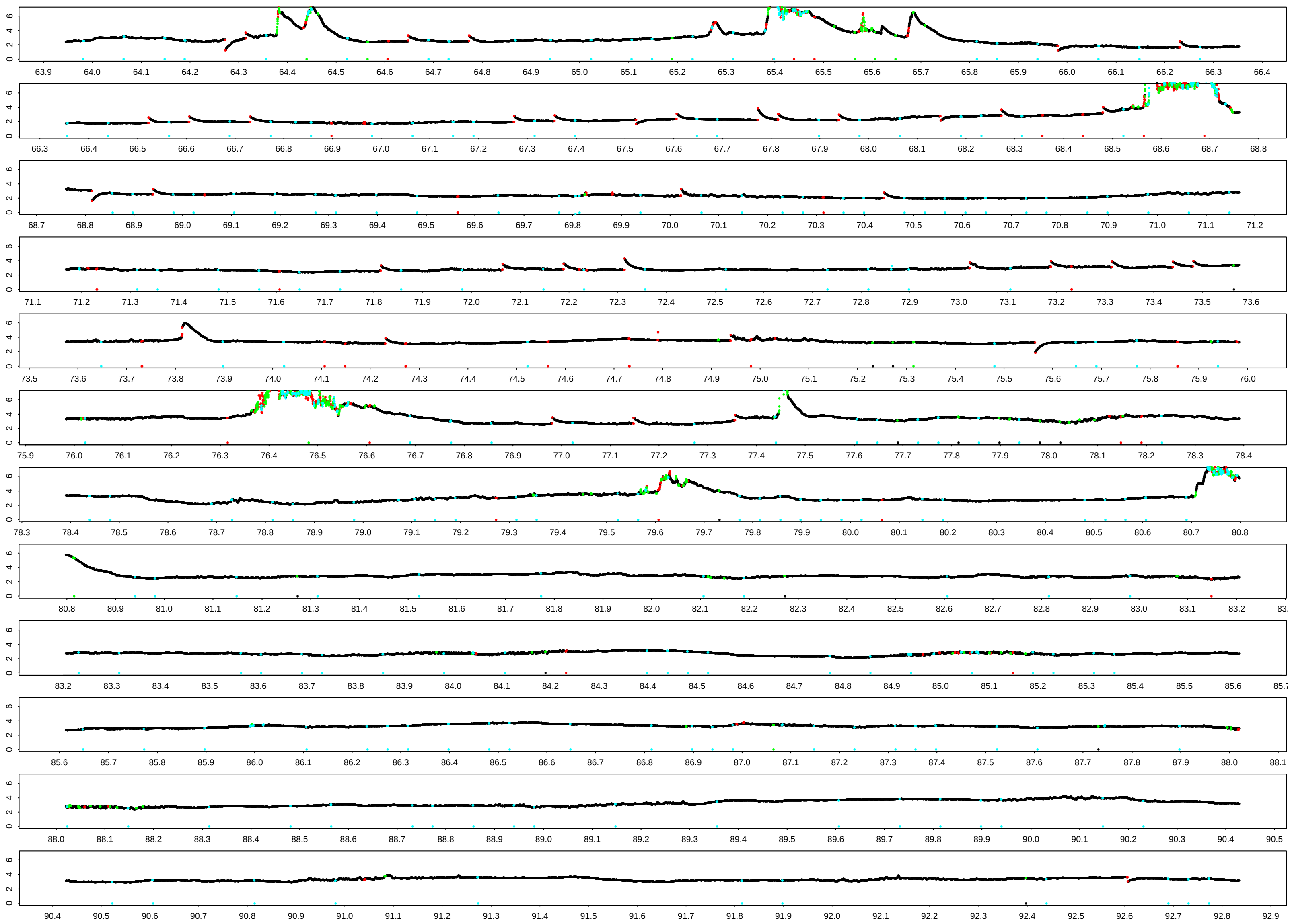

Figure 15: ARM time series segment (days vs. cm) with perturbations identified by a Kalman filter model. Perturbation free portions are black. Perturbations are in three colors: outliers cyan, slope changes red, and level changes green.

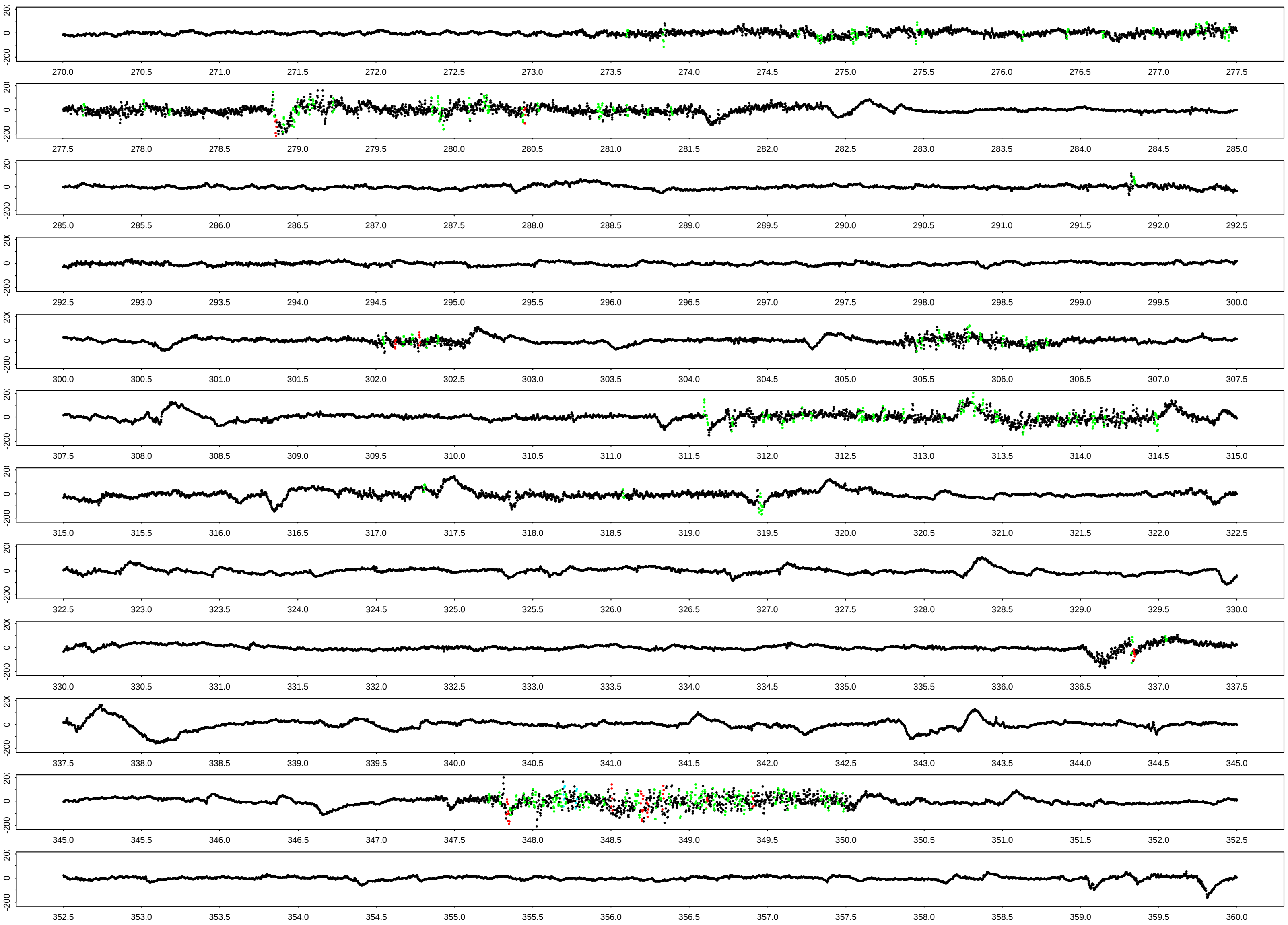

Figure 16: EEG time series segment with perturbations identified by a Kalman filter model. Perturbation free portions are black. Perturbations are in three colors: outliers cyan, slope changes red, level changes green.

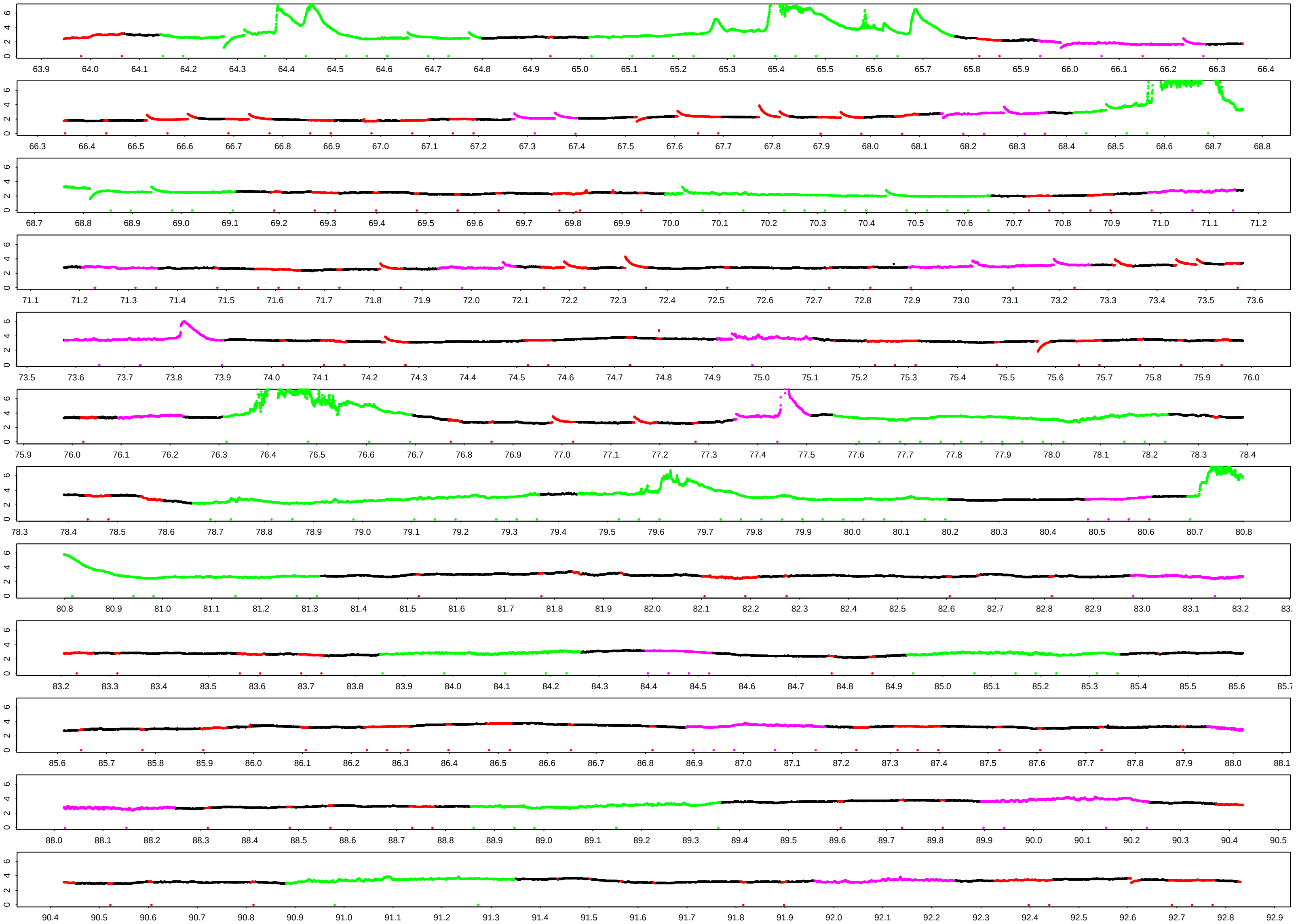

Figure 17: ARM time series segment (days vs. cm) with perturbations identified by a nonlinear dynamical model. Perturbation free portions are black. Perturbations are in three groups by duration and denoted by red, magenta, and green.

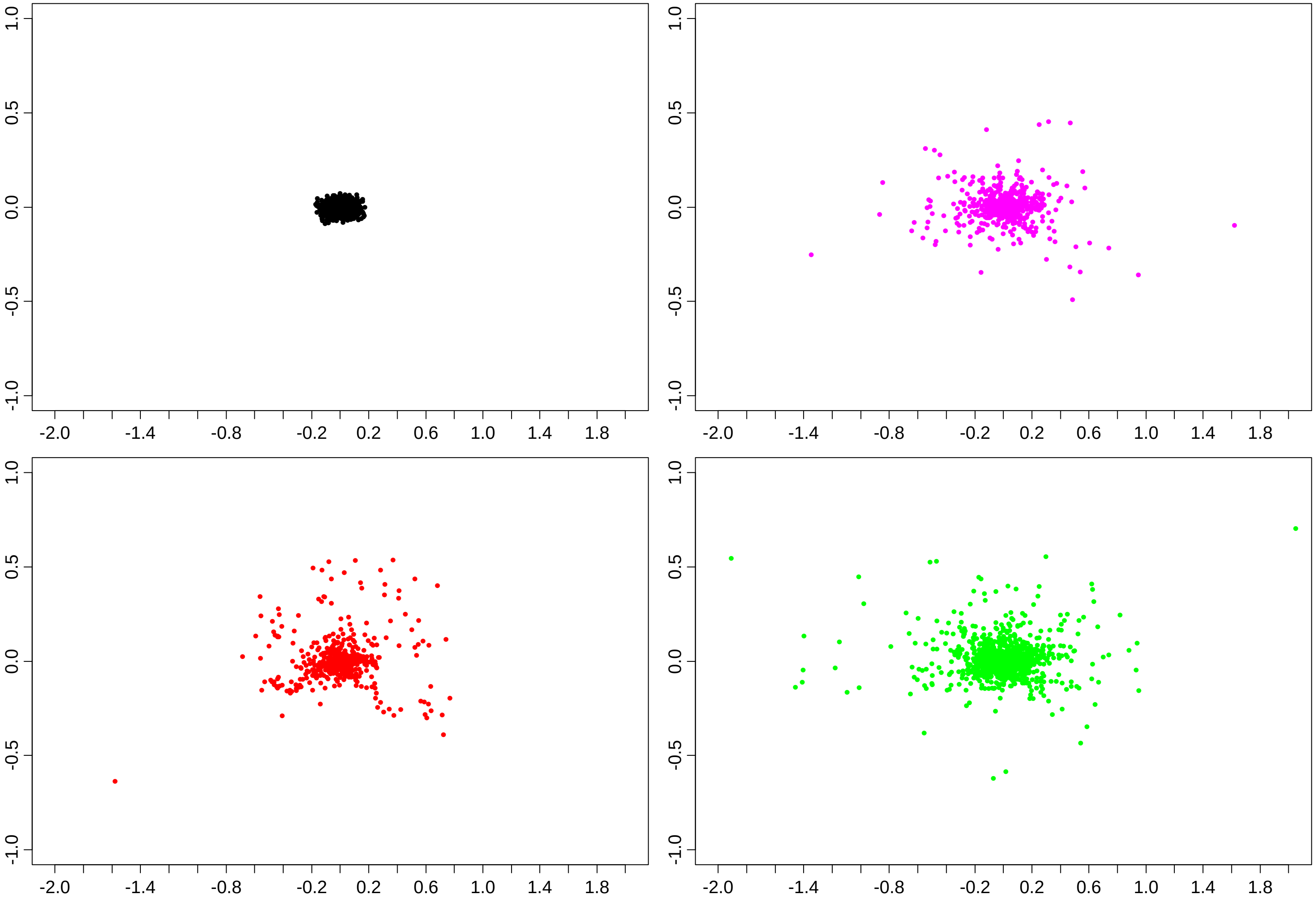

Figure 18: Scatterplots from the nonlinear dynamical model analysis of ARM time series segment. Points inside perturbation free portions are black. Perturbation points are in three groups by perturbation duration and denoted by red, magenta, and green.

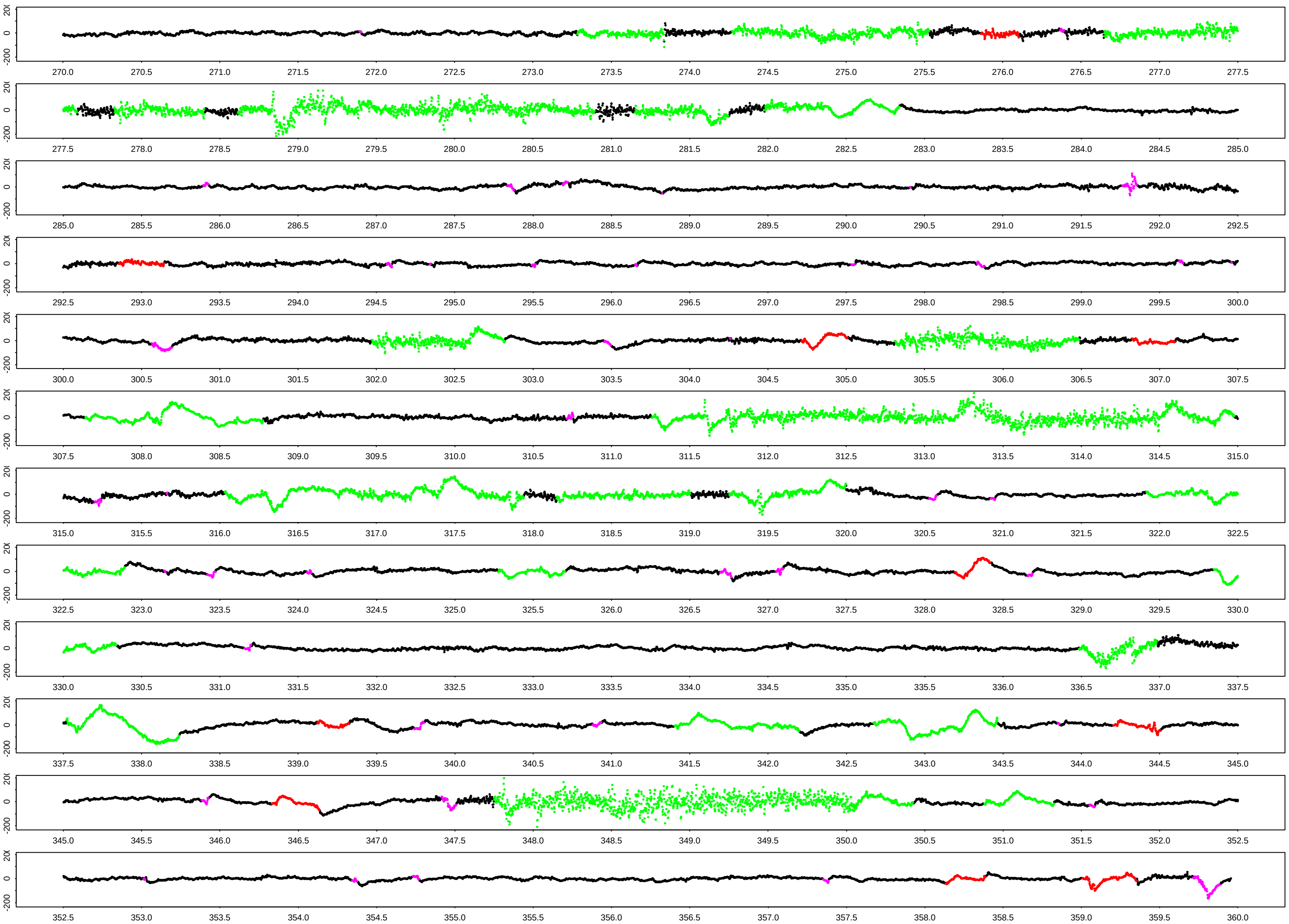

Figure 19: EEG time series segment with perturbations identified by a nonlinear dynamical model. Perturbation free portions are black. Perturbations are in three groups by duration and denoted by red, magenta, and green.

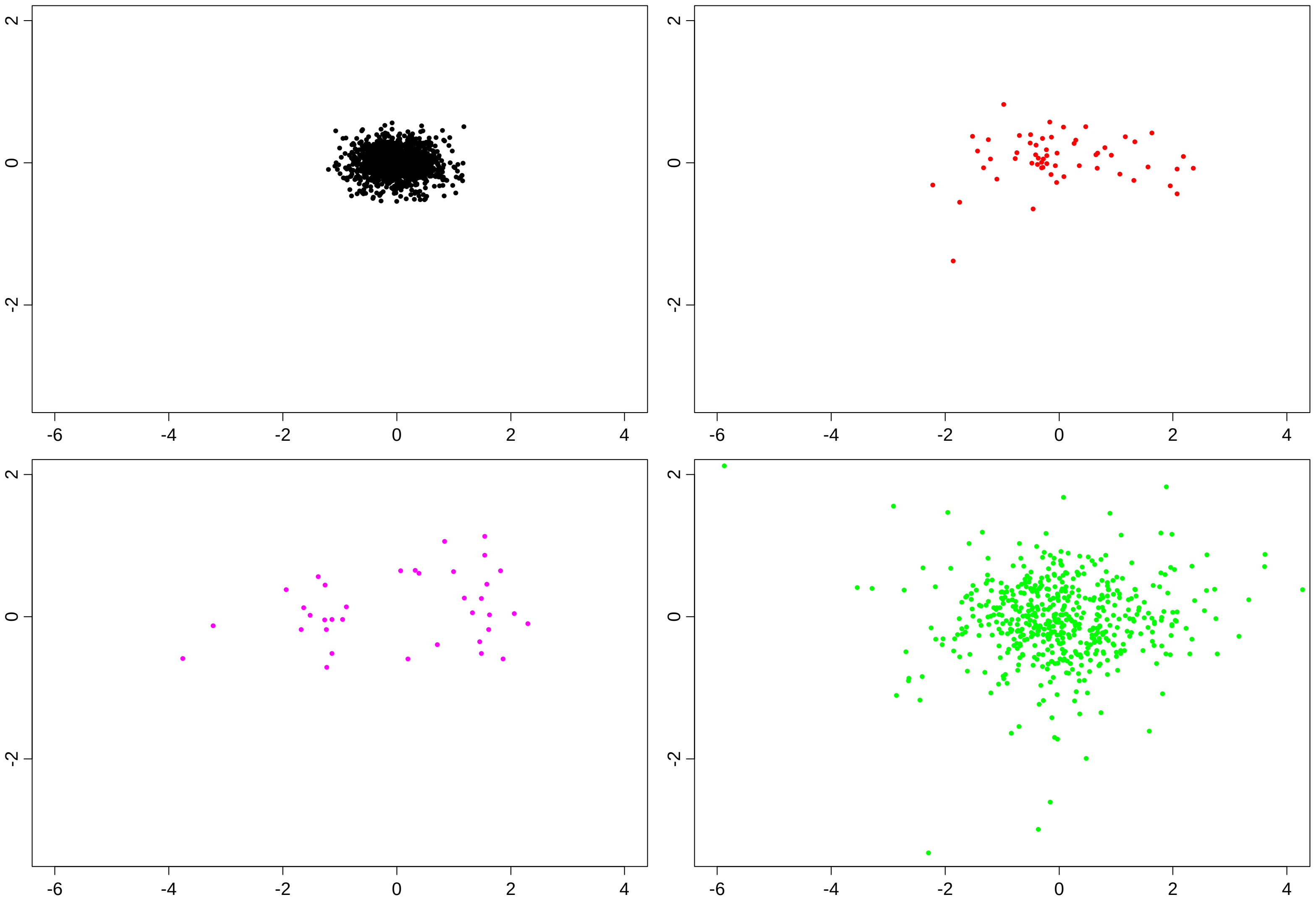

Figure 20: Scatterplots from the nonlinear dynamical model analysis of EEG time series segment. Points inside perturbation free portions are black. Perturbation points are in three groups by perturbation duration and denoted by red, magenta, and green.

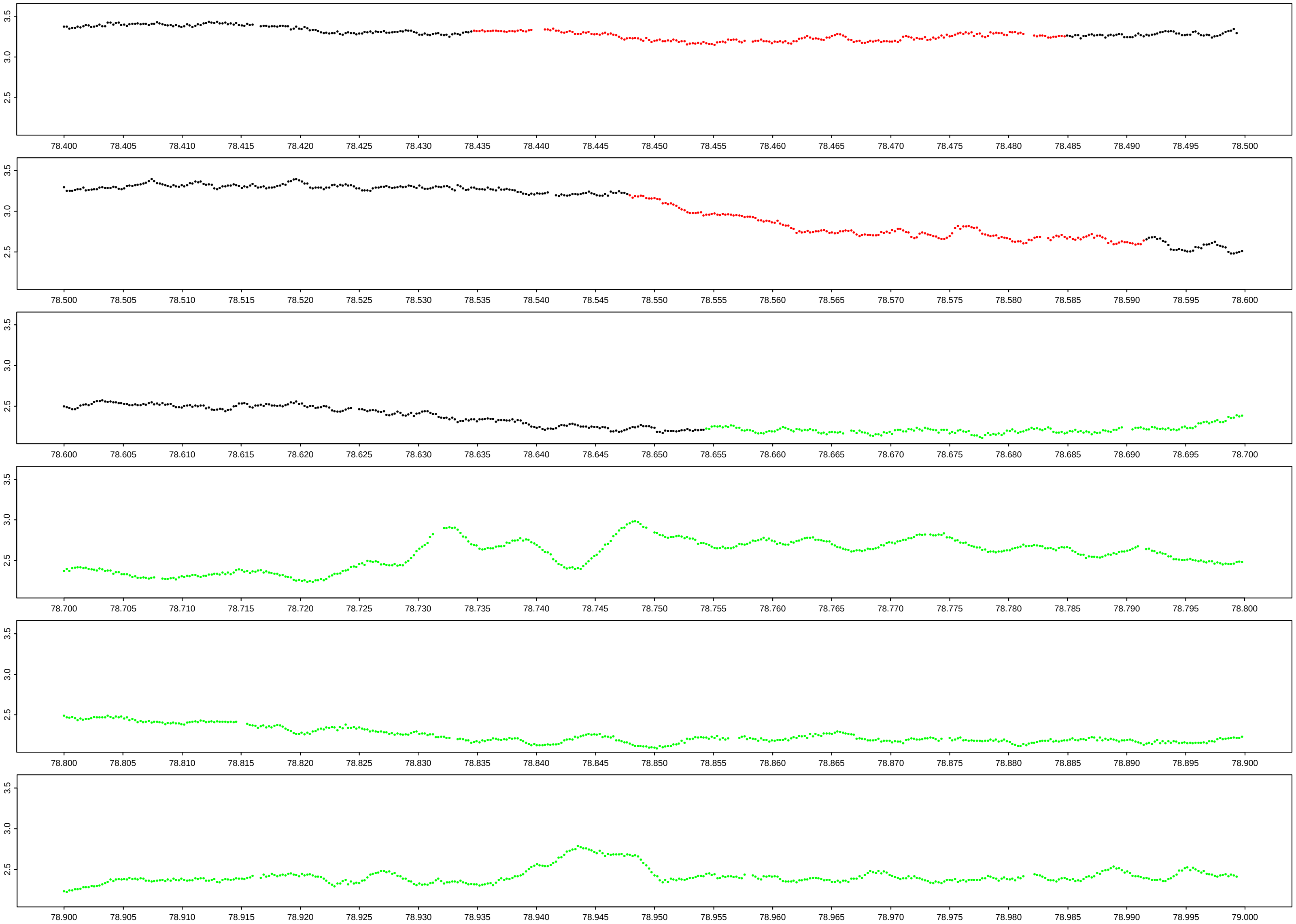

Figure 21: A magnified view of the coded ARM time series for days 78.4 through 79.0 in Fig. 17

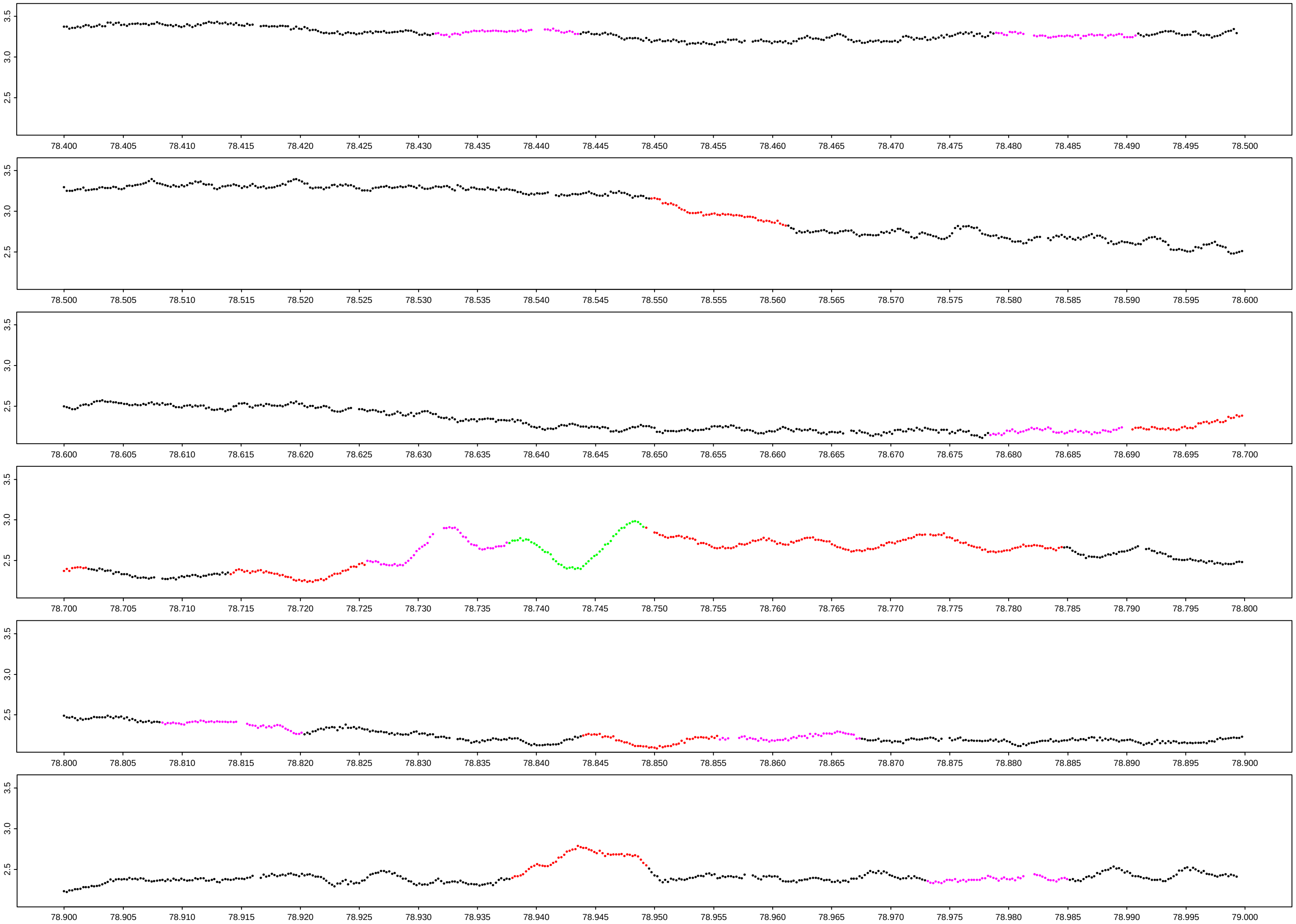

Figure 22: A magnified view of the coded ARM time series for days 78.4 through 79.0 in Fig. 9

ORNL-13114

INTERNAL DISTRIBUTION

1. M. A. Akerman
2. C. K. Bayne
3. M. D. Cheng
4. N. E. Clapp,Jr.
5. T. S. Darland
6. C. S. Daw
7–11. D. J. Downing
12. J. B. Drake
13–17. V. Fedorov
18. L. J. Gray
19. L. M. Hively
20. P. Kanciruk
21–25. W. F. Lawkins
26. M. R. Leuze
27–31. M. D. Morris
32. C. E. Oliver
33–37. G. Ostrouchov
38. V. Protopopescu
39–43. S. A. Raby
44. R. F. Sincovec
45. P. T. Singley
46. D. A. Wolf
47. K-25 Applied Tech. Library
48. Y-12 Technical Library
49. Laboratory Records - RC
50–51. Laboratory Records Dept.
52. Central Research Library
53. M&C Records Office
54. ORNL Patent Office

EXTERNAL DISTRIBUTION

55. Dr. Henry Abarbanel, Institute for Nonlinear Science, University of California, San Diego, La Jolla, CA 92093-0402

56. Prof. Cor M. van den Bleek, Dept. of Chemical Process Technology, Delft University of Technology, P.O. Box 5045, 2600 GA Delft, THE NETHERLANDS

57. Dr. Richard Beckman, Statistics Group Al, Los Alamos National Laboratory, MS F600, Los Alamos, NM 87545

58. Dr. Michael Eisenstadt, Knoxville Neurology Clinic, 930 Emerald Avenue, Suite 815, Knoxville, TN 37917

59. Dr. Celso Grebogi, Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742

60. Dr. Dan Hitchcock, Office of Scientific Computing, ER-7, Applied Mathematical Sciences, Office of Energy Research, U. S. Department of Energy, Washington, DC 20585

61. Dr. Fred Howes, Office of Scientific Computing, ER-7, Applied Mathematical Sciences, Office of Energy Research, U. S. Department of Energy, Washington, DC 20585

62. Prof. Mark Johnson, Department of Statistics, University of Central Florida, Orlando, FL 32816-0370

63. Prof. Sallie Keller-McNulty, Kansas State University, Department of Statistics, Dickins Hall, Manhattan, KS 66506

64. Prof. W. J. Kennedy, Department of Statistics, 117 Snedecor Bulding, Iowa State University, Ames, IA 50011

65. Dr. A. M. Liebetrau, Computational Sciences Department, Battelle-Northwest, P. O. Box 999, Richland, WA 99352

66. Dr. J. Liljegren, Pacific Northwest Laboratories, P. O. Box 999, Richland, WA 99352

67. Dr. Michael McKay, Statistics Group Al, Los Alamos National Laboratory, MS F600, Los Alamos, NM 87545

68. Dr. David Nelson, Director of Scientific Computing, ER-7, Applied Mathematical Sciences, Office of Energy Research, U. S. Department of Energy, Washington, DC 20585

69. Dr. Douglas W. Nychka, Statistics Department, North Carolina State University, P.O. Box 8203, Raleigh, NC 27695-8203

70. Mr. Brent Pulsipher, Computational Sciences Department, Battelle Northwest, P. O. Box 999, K1-86, Richland, WA 99352

71. Dr. Jerome Sacks, NISS, P. O. Box 14162, Research Triangle Park, NC 27709-4162

72. Dr. Jaap C. Schouten, Dept. of Chemical Process Technology, Delft University of Technology, P.O. Box 5045, 2600 GA Delft, THE NETHERLANDS

73. Prof. A. F. Smith, Department of Mathematics, University of Nottingham, University Park, Nottingham NG7 2RD, England

74. Dr. Alan Solomon, P. O. Box 227, Omer 84965, Israel

75. Dr. Daniel L. Solomon, Department of Statistics, North Carolina State University, P. O. Box 5457, Raleigh, NC 27650

76. Dr. William Thomas, University of Tennessee Veterinary School, Department of Small Animal Clinical Sciences, P. O. Box 1071, Knoxville, TN 37901-1071

77. Office of Assistant Manager for Energy Research and Development, Deaprtment of Energy, Oak Ridge Operations Office, P. O. Box 2001, Oak Ridge, TN 37831-8600

78–79. Office of Scientific and Technical Information, P. O. Box 62, Oak Ridge, TN 37830