Effective strategies for segmenting data into coherent subsets

Kevin Bleakley, Marc Lavielle

Effective strategies for segmenting data into coherent subsets

Kevin Bleakley & Marc Lavielle

November 18, 2011

Abstract

Automatic segmentation of data into coherent subsets is important in applications as varied as signal processing, bioinformatics and pharmacology. Under this general framework, we investigate the problem of data-driven reconstruction of an unknown, piecewise-constant density function and propose two methods to solve it; the first is directly inspired by the segmentation approach, whereas the second uses a maximum likelihood approach. Motivated by a problem in pharmacometrics, we then introduce a segmentation algorithm which fits into the same general framework and is used for automatically binning data for model assessment purposes.

Keywords: Segmentation; Signal; Piecewise-constant density function; Maximum likelihood; Pharmacology; Visual Predictive Check; Dynamic programming; Histogram.

1 Introduction

In numerous real-world situations it is desirable to take a set of data and break it up into subsets which are meaningful for the given context. In a sense, the context induces knowledge as to *what a typical subset looks like*, and the question becomes to encode this knowledge into a well-defined mathematical criteria that can automatically create relevant subsets.

If data is multi-dimensional, a natural framework for creating subsets is *clustering* (see [1] for review) whereby some function of the coordinates/features of each object is used to define clusters of data. A simple

example is to separate a mixed bag of oranges and plums into two clusters based on the features *colour* and *diameter*.

One type of clustering is of particular interest for the present article. It involves cases where one data coordinate/feature has a natural ordering along an axis (e.g., time) and a constraint is imposed that clusters must be sets of consecutive data points with respect to this feature. This is known as *segmentation*.

Segmentation is a pervasive goal in several distinct settings. One main body of research and application is signal processing. For instance, one may want to identify speech segments from a radio broadcast or segment an audio stream into acoustically homogenous blocks [2, 3]. More generally, one may wish to detect the set of points at which a signal's variance changes, thus dividing the data points into "constant variance" segments [4].

In the field of biology, segmentation of genomic profiles is also a commonly performed task. The most well-known case is the analysis of copy number profiles, which present the ratio of DNA from diseased cells compared with normal cells at a set of *ordered* points along the genome. Contiguous regions of a disease genome can be copy-number normal (ratio = 1), gained (ratio > 1) or lost (ratio < 1), and the goal is to demarcate the boundary between regions of different constant copy number. A large number of solutions have been proposed to do this [5, 6, 7, 8, 9, 10, 11].

Several of these signal segmentation strategies share a certain methodology that we would like to generalise to settings where the data is not a signal in the traditional sense. To explain, suppose that our n data points have a feature x that is (or can be) ordered: $x_1 \leq x_2 \leq \ldots \leq x_n$. Furthermore, suppose that we have a real-valued function g so that for any *partition* $I = \{I_1, \ldots, I_K\}$ of the indices $1, \ldots, n$ into K distinct subsets $I_k = \{i_{k-1} + 1, \ldots, i_k\}, k = 1, \ldots, K$ (with by definition $i_0 = 0$ and $i_K = n$), we can calculate $g(\{x_i, i \in I_k\})$. This function g gives a measure of how well the subset matches what a *typical subset should look like* for a given context. Then, we can propose to minimise

$$J(I) = \sum_{k=1}^{K} g(\{x_i, i \in I_k\})$$

over *all* possible partitions I of size K.

For a given K, such a minimisation is intractable via brute force (i.e., testing every possible partition) once $n >> K$, as there are $\binom{n-1}{K-1}$ possibilities.

However, using dynamic programming (see [12] for review), it turns out that it suffices to pre-calculate g for each possible segment (which can be done in $\mathcal{O}(n^2)$ operations rather than $\mathcal{O}(n^K)$), then run an iterative algorithm (also $\mathcal{O}(n^2)$) to find the minimum for a given K.

In this paper, we therefore consider problems which can be reformulated as segmentation tasks, and then attempt to solve them by defining a suitable function g and running a dynamic programming algorithm to find optimal solutions. The unifying theme is that the choice of a relevant and computationally feasible g is a critical step in marrying each situation to a intuitively reasonable segmentation.

In the first part of the paper, we attempt to answer the following question: given data generated from an unknown piecewise-constant density function (i.e., a density that looks like a histogram), recover the density function; that is, locate the set of points at which the density changes, and thus the shape of the histogram. We frame the search for the set of (change-) points as a 1-d segmentation along the x axis; the boundaries between segments are the segmentation points we are looking for. We propose two methods to do this, each of which requires recourse to the definition of a function g as well as dynamic programming.

In the second part of the paper, we consider a more general problem, that of constructing a good histogram representation of a 1-d finite data set in any chosen subject of interest, whether it be mathematics, psychology, biology, marketing, etc. A "good" histogram should aim to be made up of a set of bins which give a reasonable "discrete summary" of data which itself may be discrete or continuous. The two simplest binning strategies are: i) *equal-width* (k bins of the same width); and ii) *equal-size* (same number of data points in each bin). However, for a finite data set, both can give poor summaries of the data. For example, if there is a high density of points in a small localised region that is surround by low density, both methods can lead to putting a bin edge in the middle of the high density area, thus mixing high and low density regions in bins to the left and right, giving a poor local summary of the data. Variable-width histograms, like those we will propose in the present paper using a segmentation-based approach, can improve the histogram "summary" of finite data sets.

Before introducing this approach, let us describe the application which provides the motivation for the second part of the paper. A Visual Predictive Check (VPC) is a graphical tool for model assessment used to compare the distribution of real observations with that of simulated data [13, 14, 15, 16].

Summary statistics of the observed and simulated data are compared *visually*. The simulated data itself is generated from the mathematical model expected to characterise the underlying biological process. Typically, the summary statistics are related to the median and two extreme percentiles, e.g., the 10^{th} and 90^{th}. Thus, for time-course data one can plot the relevant median and percentiles of both the real and simulated data with respect to time, and visually compare them. If the model is good, we would expect the simulated median and percentiles to be systematically "close" to the real data ones. One strategy to help quantify this is to create a confidence interval (CI) for the percentiles based on the simulated data, and then visually check how well the percentiles calculated on the real data "fit inside" the interval [17].

When trying to visually compare the real and simulated data as above, the real data are first typically binned into specific time intervals; otherwise, the predicted CIs may exhibit overly "bumpy" patterns, making visual interpretation difficult. As mentioned earlier, there are two "simple" binning strategies: *equal-width* and *equal-size*, that apply to pharmacometric time-course data. Unfortunately, the design of typical experiments makes both these options inherently poor "summaries" of the real data. This may end up hiding the evidence of a poor model choice, or incorrectly rejecting the correct model when doing a VPC.

In this contribution, we present a binning strategy tuned to pharmacometric time-course data that automatically determines a "good" binning, i.e., a well-chosen *number of bins* and their *edges*. Note that in this application, we are not building a histogram *per se*, merely finding the edges of the data bins; but the methodology leads directly to building a histogram if that is the final goal. As before, the binning strategy we introduce involves a suitably defined function g and dynamic programming to determine the edges, as well as a model-selection approach to select the number of bins. In practice, this leads to irregularly sized bins that better correspond to the clusters we see in the data. Consequently, we improve the match between the real data and the VPC "summary", leading to better model diagnosis in practice.

2 Data-driven reconstruction of piecewise-constant density functions

Suppose that we have data generated from an unknown piecewise-constant density function

$$f(x) = \sum_{k=1}^{K} \frac{p_k}{z_k - z_{k-1}} \mathbf{1}_{\{z_{k-1} \leq x < z_k\}} \tag{1}$$

on $[0, 1]$, where $0 = z_0 < z_1 < \ldots < z_K = 1$ are the bin boundaries and p_k the probability that x is contained in the kth bin. We have that $\sum p_k = 1$, and for all purposes, f is a density that looks like a histogram. We are interested in trying to reconstruct f using data $X_i \sim f$. In the following, we present two ways to attack the problem. The first, inspired by the statistical framework for finding change-points in (or segmenting) signals, is to consider the data to be a *signal* on $[0, 1]$ and to look for places where this signal has change-points; this means finding the location of the edges of the piecewise-constant bins. The second is a maximum-likelihood approach which predicts the location of bin edges. The idea which unifies these two methods, along with the VPC method to be described subsequently, is that in each case, we can define a suitable function g and a quantity $J(I)$ that can be optimised over all partitions I of the data into K subsets.

2.1 A change-point estimation approach

Suppose we have $X_i \sim f$, $i = 1, \ldots, n$ where f is a piecewise-constant density function on $[0, 1]$. As the density is uniform in each bin, the inverse F^{-1} of the true cumulative distribution function (cdf) F is piecewise-linear and continuous, and as long as we never have $p_k = p_{k+1}$, the slope of F^{-1} changes at each bin edge. Using data $x_1, \ldots, x_n$, we can construct the inverse of the empirical cdf F_n^{-1} (plotted as $\{(i/n, x_i), i = 1, \ldots, n\}$) and we know that by Glivenko-Cantolli, F_n^{-1} converges uniformly to F^{-1}. This motivates the following method.

For a given number of bins K, we propose to construct a linear spline with knots $\tau_1, \ldots, \tau_{K-1}$ to approximate the unknown F^{-1}. If the knot locations are allowed to be in general position, there is no analytic solution to this problem, and care must be taken to avoid local minima; a great deal of research has been undertaken to provide reasonable solutions (see [18] for

review). We therefore limit ourselves to splines with knots located at $u_i := \frac{i}{n}$. We propose to estimate F^{-1} by minimising the criteria

$$J_{\mathrm{LS}}(I) := \sum_{k=1}^{K} \sum_{i \in I_k} (x_i - a_k u_i - b_k)^2 \tag{2}$$

over all partitions I with K subsets, with the spline continuity constraint $a_k \tau_k + b_k = a_{k+1} \tau_k + b_{k+1}$ for $k = 1, \ldots, K-1$.

2.1.1 Algorithm

An exhaustive search for the solution of this minimisation problem remains intractable for even moderately sized n as we must search among $\binom{n-1}{K-1}$ solutions.

We remark that without any continuity constraint, the criteria J_{LS} defined in (2) decomposes as a sum of independent terms over each bin. Then, minimisation of J_{LS} is straightforward using a dynamic programming algorithm [12]. Unfortunately, the spline continuity constraint means that we cannot directly do this. A practical alternative consists of: 1) minimising $J_{\mathrm{LS}}(I)$ without the continuity constraint, using dynamic programming; 2) placing a knot at the $K-1$ end-points $\frac{i_1}{n}, \ldots, \frac{i_{K-1}}{n}$ of each subset I_k of the optimal partition I^*; 3) attempting to improve the discontinuous solution with a continuous one. Precisely, we independently and randomly move each knot in order to improve the criteria, and repeat until convergence.

We deliberately do not go into further details about this procedure since the alternative proposed below will be shown to be much easier to implement while providing better results.

2.2 A maximum-likelihood approach

We now propose an alternative approach based on parameter estimation which turns out to have good computational properties and, as will be seen, generally superior performance. Suppose once more that the density is given by (1). For data $x_1, \ldots, x_n$ generated from f, it is straightforward to calculate the relevant log-likelihood:

$$\mathcal{L}(x, p, z) = -\sum_{k=1}^{K} \sum_{i=1}^{n} \mathbf{1}_{\{z_{k-1} \leq x_i < z_k\}} \log \left(\frac{z_k - z_{k-1}}{p_k} \right).$$

Maximising with respect to p and substituting back in, we obtain:

$$\mathcal{L}(x, p(z), z) = -\sum_{k=1}^{K}\sum_{i=1}^{n} \mathbf{1}_{\{z_{k-1}\leq x_i < z_k\}} \log\left(\frac{n(z_k - z_{k-1})}{\sum_{i=1}^{n} \mathbf{1}_{\{z_{k-1}\leq x_i < z_k\}}}\right). \quad (3)$$

It remains to provide data-driven estimates of the z_k. For any partition I of the data into K subsets, let us define $n_k := i_k - i_{k-1}$ and $d_k := x_{i_k} - x_{i_{k-1}}$ (with convention $x_0 = 0$ and $x_K = 1$). Estimating z_k with x_{i_k} in (3) would imply that we should try to minimize the criteria

$$\begin{aligned} J_{\mathrm{ML}}(I) &:= -2\,\mathcal{L}(x, p(z), z) & (4)\\ &= 2\sum_{k=1}^{K} n_k \log\left(\frac{n\,d_k}{n_k}\right) & (5) \end{aligned}$$

over all partitions I of size K. Moreover, J_{ML} is a criteria like J_{LS} that can be optimised over all partitions of the data into K subsets using dynamic programming.

Remark: Denoting $(\hat{z}_k)$ the estimates of the (z_k), we can show using the general results of [19] that $\max_k \|\hat{z}_k - z_k\| = \mathcal{O}_p(1/n)$.

2.3 A comparison of the two methods

Figures 1(a)-(c) illustrate the first change-point estimation approach for one simulated trial with $n = 100$ data points. The original piecewise-constant density is displayed in Figure 1(a). Figure 1(b) shows the inverse of the original cumulative density function F^{-1} together with the observations. Here, F^{-1} is a continuous and piecewise-linear function. The knots are the limits of the bins. Figure 1(c) displays the estimate of F^{-1} obtained with the first method, i.e., by minimising criteria J_{LS} under the continuity constraint. We see that the obtained solution fits very well the observed data but the locations of the estimated bins slightly differ from the original ones.

Figures 2(a)-(c) compare the original density function with various estimates. Figure 2(a) displays the optimal empirical estimate, assuming that the location of the bins is known. For a given sequence of data, this is clearly the best estimate that we can expect to obtain. The second estimate, displayed in Figure 2(b), is the least-square estimate obtained by minimising J_{LS} under the continuity constraint. It is therefore directly derived from the

estimate of F^{-1} displayed Figure 1(c). Lastly, Figure 2(c) presents the maximum likelihood estimate, obtained by minimising J_{ML}. We remark that for this particular example, the maximum likelihood solution almost coincides with the optimal empirical estimate.

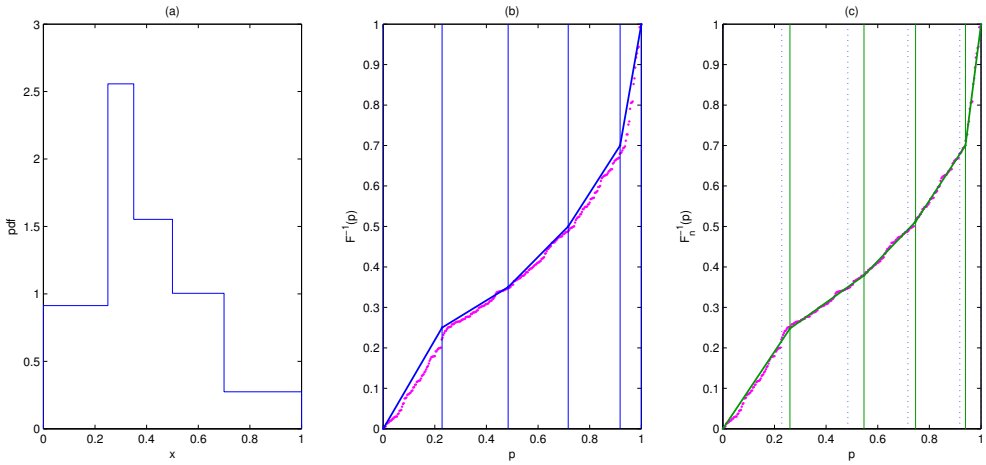

Figure 1: **Change-point estimation for density reconstruction.** (a) The original 5-bin histogram-like density f; (b) its inverse cumulative distribution function (cdf) F^{-1} in blue and the observations in magenta. Vertical lines indicate the change-points, *i.e.* the limits of the bins; (c) the least-squares optimal 5-segment piecewise-linear smoothing obtained by minimising J_{LS} under the continuity constraint is displayed in green. Green (resp. dotted) vertical lines indicate the estimated (resp. original) change-points.

Obviously a unique simulation is not enough for demonstrating the superiority of a given method. Thus, in 100 trials, we generated $n \in \{20, 50, 100, 200, 500, 1000\}$ data points from random 5-bin histogram-like densites on $[0, 1]$, then applied the two algorithms to attempt to reconstruct the true densities. Reconstruction quality was measured as the average distance (i.e., integral of the difference) between the true and estimated densities (as in Figure 2). Results were averaged over the 100 trials and are presented in Figure 3 for both the change-point and maximum-likelihood approaches. We see clearly that the maximum likelihood approach gives better results than the least-square approach, for any sample size. Futhermore, for

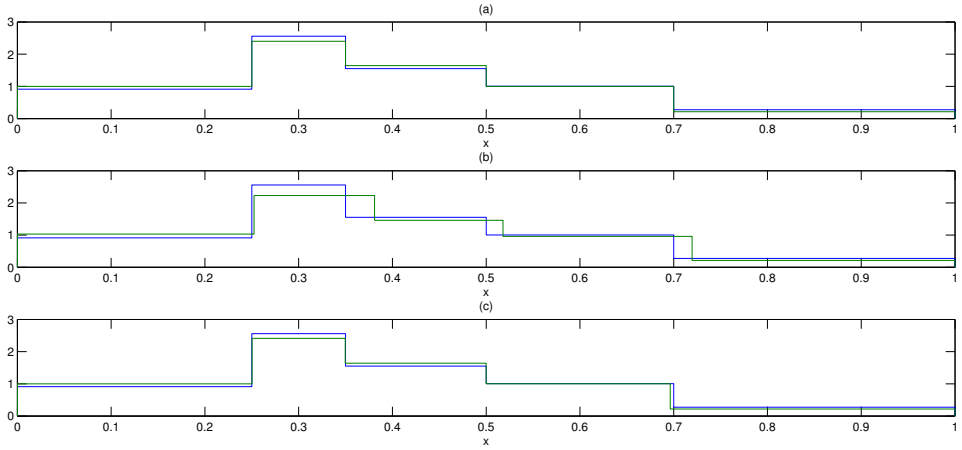

Figure 2: **Comparison of different estimators for density reconstruction.** (a) The original 5-bin histogram-like density f in blue and the optimal empirical estimate in green, assuming that the location of the bins is known, (b) the least-squares estimate obtained by minimising J_{LS} under the continuity constraint, (c) the maximum likelihood estimate, obtained by minimising J_{ML}.

$n = 1000$, the maximum likelihood approach is nearly 300 times faster (0.13 secs compared with 36 secs) on a typical current laptop computer. Lastly, as described above, the maximum likelihood estimate can easily be computed using a dynamic programming algorithm. We therefore strongly recommend this method for reconstructing piecewise constant densities, and consider only it for model selection below.

2.4 A model selection approach for estimating the number of bins

When the number of bins is unknown, we propose to estimate it by minimising a penalised criteria:

$$U(I) = J_{\mathrm{ML}}(I) + \mathrm{pen}(K(I)), \tag{6}$$

where $K(I)$ is the number of bins of partition I and 'pen' is an increasing function.

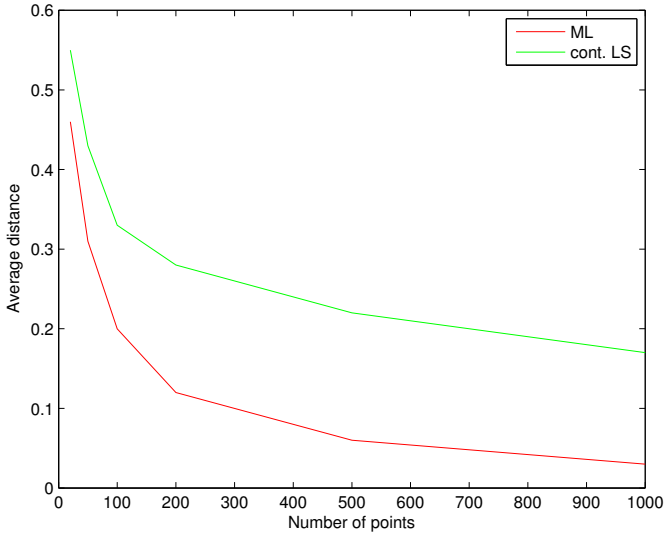

Figure 3: Reconstruction quality as a function of the number of data points for the continuous version of the change-point estimation (cont. LS) and the maximum-likelihood approach (ML).

The unknown criteria that we wish to optimise using this approach is the expectation $E\left(d(f_n^\star, \hat{f}_n)\right)$ of the distance between the estimated density $\hat{f}_n$, obtained with the proposed method, and the optimal empirical density $f_n^\star$, obtained using the original knots. This expectation is estimated by Monte-Carlo: we perform a very large number of simulations ($N = 16000$ here) with different sample sizes (between $n = 200$ and $n = 2000$) and different density functions f. We then test different penalisations for which we estimate the expectation of this criteria from the N simulations and look for the penalisation which minimises this expectation.

The Bayesian Information Criteria (BIC) suggests defining $\text{pen}(K) = 2 * \log(n) * K$ since there are 2 parameters per bin to estimate (the density and the right limit). Extensive numerical experiments we performed showed that BIC provided poor results for estimating the number of bins and thus the unknown density function. Instead, these experiments suggested the use

of a penalisation of the form

$$\text{pen}(K) = K * (\beta_1 + \beta_2 \log(n)). \tag{7}$$

For a given density f, observations and values of β_1 and β_2, let $\hat{f}_n(\beta_1, \beta_2)$ be the estimated density obtained by minimising (7). We obtained optimal values of $\beta_1 = -15$ and $\beta_2 = 3.7$ by calibration, *i.e.* by minimising the criteria $E\left(d(f_n^\star, \hat{f}_n(\beta_1, \beta_2))\right)$ estimated from the N simulations.

Remark 1: Using the general results of [19], we can show that the estimate $\hat{K}_n$ of the number of bins converges to the true number K.

Remark 2: In a non-asymptotic context, Birgé and Massart [25] have shown that a penalty function of the form

$$\text{pen}(K) = K * (\beta_1 + \beta_2 \log(n/K))$$

is optimal for recovering a piecewise-constant signal. Here, using $\log(n/K)$ instead of $\log(n)$ does not improve the results.

Remark 3: We did not succeed in improving this penalisation approach using several adaptive procedures for model selection. The main idea of these adaptive methods is to assume that the criteria J_{ML} is a linear function of the penalisation for models of high dimension (i.e., models with more bins than the original partition) and detect from which dimension this linear assumption starts to be valid [4, 25]. The slope heuristic proposed by Birgé and Massart in [25] was quite unstable and could not be implemented properly in this context. The method proposed by Lavielle in [4] for detecting an abrupt change in the second derivative of the criteria gave good results, but on average less good than the penalisation approach.

Remark 4: The risk ratio measures how well the selected partition performs in comparison to the oracle, i.e., the optimal partition which minimises the distance $d(f_n^\star, \hat{f}_n)$. Here, the risk ratio can be estimated by Monte-Carlo using a new set of simulated data. The risk ratio is 1.27 with the proposed penalisation method. It is 1.54 and 1.85 for BIC and the Akaike criteria, respectively, and 1.40 with the adaptive procedure proposed by Lavielle in [4]. If the number of bins is fixed to the true number of bins, then the risk ratio is 1.19.

3 Data segmentation for visual predictive checks

3.1 What are visual predictive checks?

Visual predictive checks, or VPCs, are model evaluation methods for evaluating stochastic models [13, 14, 15, 16]. They provide a way to help decide whether a model correctly describes given data, as well as to decide if the model is likely to predict responses in future subjects.

To perform a VPC, first several sets of data are simulated with the proposed model. Then, the distribution of the simulated data is compared with the empirical distribution of the true data. Let us describe this VPC methodology – illustrated in Figure 4 – as implemented in MONOLIX (`www.monolix.org`), a software dedicated to the analysis of nonlinear mixed effects models: a) Observations $(y_i; 1 \leq i \leq n)$ are measured at times $(x_i; 1 \leq i \leq n)$. n is the total number of observations across the *whole set of individuals*; b) Data is grouped into adjacent time intervals (bins); c) Several empirical percentiles are computed for the data in each bin; d) A large number of datasets are simulated under the model being evaluated, using the design of the original dataset; e) The data from each simulated dataset is grouped into the same bins; f) The same percentiles are computed in each bin for each of the simulated datasets; g) Confidence intervals (CI) for each percentile are calculated using these simulated percentiles; h) Observed percentiles are compared with these CIs; i) Regions where the observed percentiles are not found within the CIs are filled in with red, in order to help detect misspecified models. Note that a small number of such regions does not necessarily mean a misspecified model; indeed, it is expected, and the modeler must make a decision as to whether there are too many such regions.

3.2 Binning

The problem of interest for the current article is to develop a binning (segmentation) algorithm that is useful in the VPC setting. Furthermore, we would like the algorithm to integrate into the same framework whereby it suffices to define an intuitive criteria to optimise over partitions of the data.

Let us suppose that the (pooled) data $X_i \in \mathbb{R}$, $i = 1, \ldots, n$ is generated following some (unknown) probability density function $X_i \sim f$. In the VPC application, X is the variable *time*. Binning the generated data, i.e., grouping

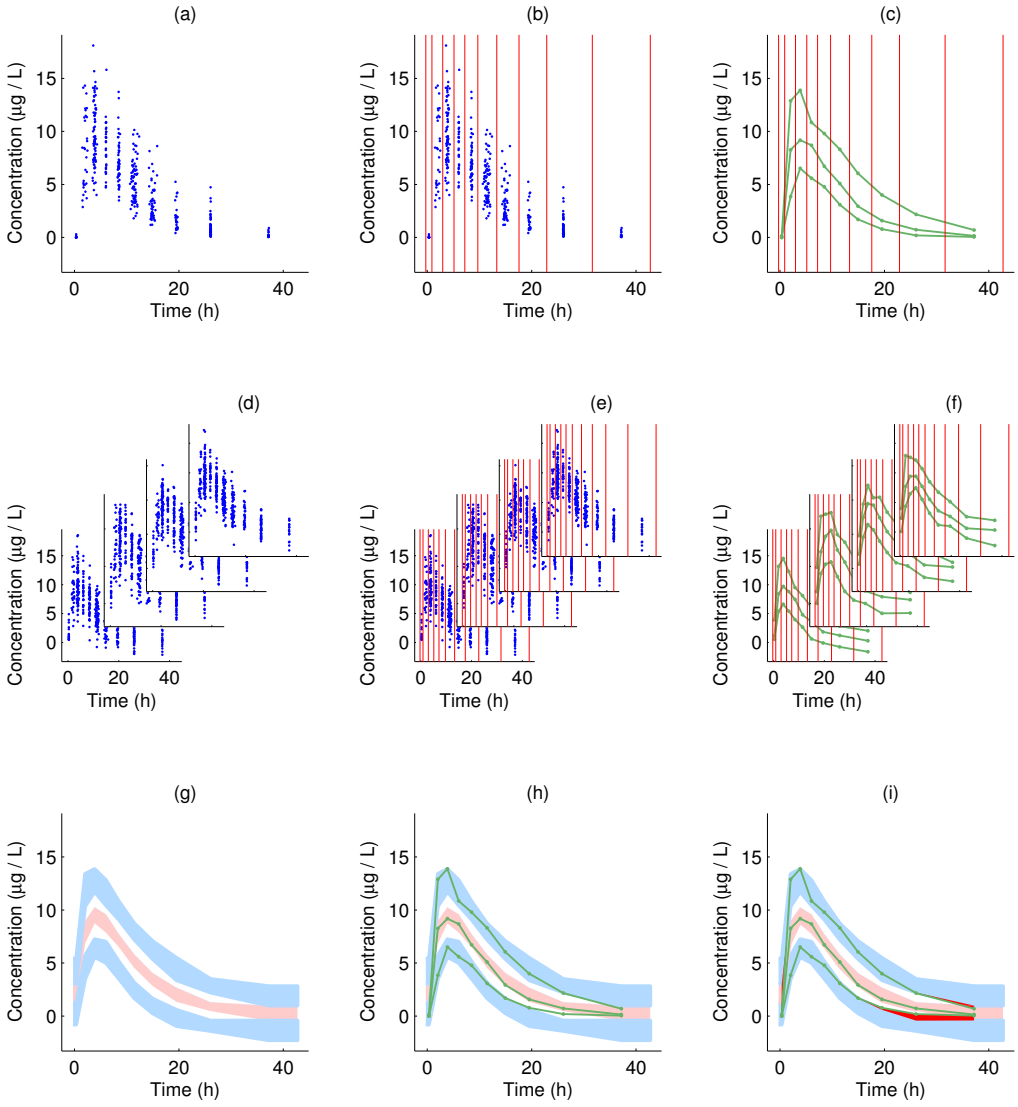

Figure 4: **Visual Predictive Check construction:** (a) the data, (b) data grouped into bins, (c) empirical 10^{th}, 50^{th} and 90^{th} percentiles computed for each bin, (d) several simulated data sets, (e) these simulated data sets grouped into the same bins, (f) the 10^{th}, 50^{th} and 90^{th} percentiles of each simulated data set computed for each bin, (g) 90% confidence intervals computed from the percentiles of the simulated data, (h) observed percentiles and 90% confidence intervals, (i) zones outside of the confidence intervals are filled in with red.

the x_i into intervals, leads to an approximation of this distribution by a piecewise-constant one. A binning strategy should aim to be "good" as in: i) for a given number of bins, the locations of the bin edges must be chosen so as to minimise heterogeneity of the data in each bin; and ii) the number of bins must be carefully chosen, i.e., we require a good tradeoff between a large number of bins and a large number of observations in each bin; the true distribution can be accurately approximated by a piecewise-constant distribution with a large number of bins, while a large number of observations in each bin is required to accurately estimate this true distribution.

3.3 Standard binning strategies

There are various ways to implement binning. The two simplest are: i) *equal-width binning*: K bins of length $(x_{max} - x_{min})/K$; and ii) *equal-size binning*: K bins, each with n/K data points. If n is not a multiple of K, we can correct so that each bin has either $[n/K]$ or $[n/K] + 1$ data points.

In practice, equal-width binning is not appropriate when time-points are inhomogenously distributed (as is often the case in the VPC application); some bins contain many data points whereas others are completely empty. Due to this inherent poor adaptability, we do not consider the method in the following.

In other situations, several observations are obtained from different patients at the same time points. This is the case for example in the warfarin pharmacokinetic (PK) data shown in Figure 5(a). This poses obvious problems for equal-size binning.

We may wonder if the equal-size binning procedure can be modified to deal with this case of identical time points, but different number of measurements at each time point? Our first objective is to propose an automatic procedure which selects bins with "similar" amounts of data in each. Let $x_1 < x_2 < \ldots < x_M$ be the M different time points and $m_1, m_2, \ldots, m_M$ the number of measurements taken at each of these time points. As before, $n = \sum m_i$ is the total number of data. For a given number K of bins, we look for the bins $I = (I_1, I_2, \ldots, I_K)$ that minimise the following criteria:

$$J_{\mathrm{ES}}(I) = \sum_{k=1}^{K} \left| \sum_{i \in I_k} m_i - \frac{n}{K} \right|. \tag{8}$$

This can be done using dynamic programming in the same way as for the two

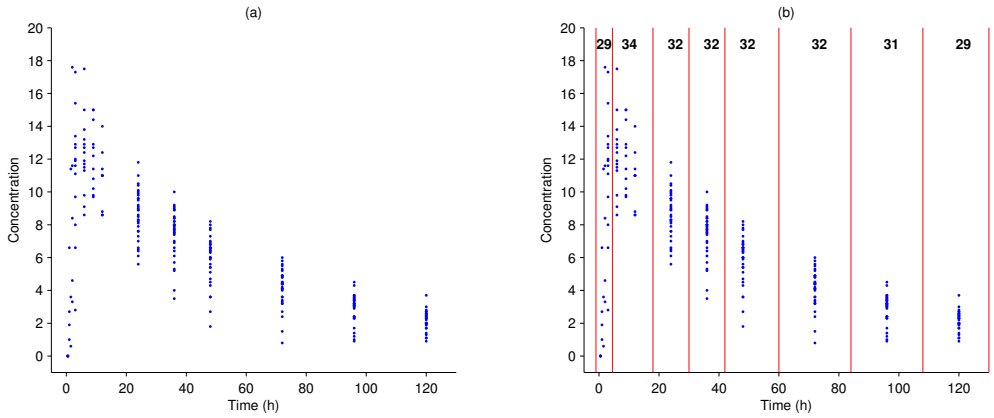

Figure 5: (a) warfarin pharmacokinetic data, (b) "approximately" equal-size binning.

previously described algorithms. The segmentation displayed in Figure 5(b) was obtained by minimising the criteria J_{ES} with $K = 8$ bins.

3.4 A new binning procedure

Often, we have data where all time points are different and the data is "clustered" around various time points, e.g., Fig. 6(a) – simulated data. In this case, the similar-size solution obtained by minimising J_{ES} no longer provides a plausible binning (Fig. 6(b)) as it does not take into account knowledge of the clusters.

One way to resolve this more general problem is to interpret binning as *clustering* or *1-d segmentation*, i.e., grouping the n time points $x_1 \leq x_2 \leq \ldots \leq x_n$ into K clusters or segments along the time axis. One possible way to do this is by *1-d* K-means clustering [24]. Let us define

$$J_{\mathrm{KM}}(I) = \sum_{k=1}^{K} \sum_{i \in I_k} (x_i - \overline{x}_k)^2 , \tag{9}$$

where $\overline{x}_k$ is the empirical mean of the x_i's in bin I_k:

$$\overline{x}_k = \frac{1}{n_k} \sum_{i \in I_k} x_i,$$

with n_k the number of points in bin k. Then, the K-means solution is found by minimising J_{KM} over all possible segmentations $I = (I_1, I_2, \ldots, I_K)$ of the data into K bins. As has become habitual, we can do this using dynamic programming [12]. Fig. 6(c) shows the optimal binning obtained by minimising J_{KM} for $K = 6$.

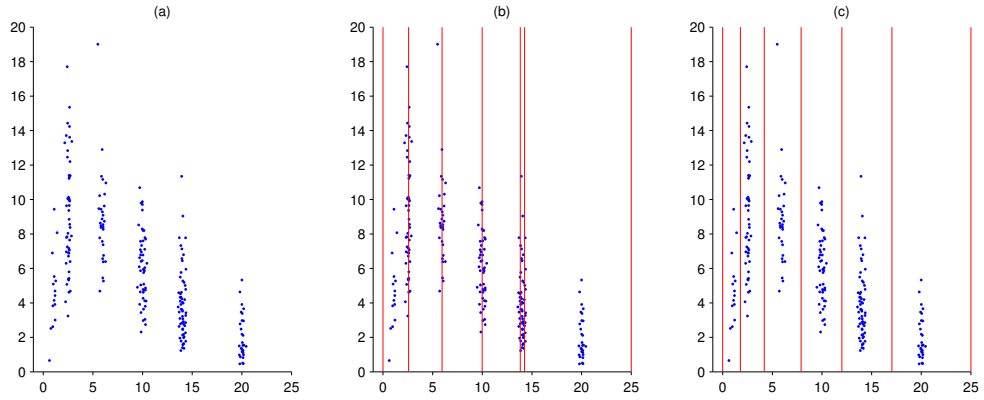

Figure 6: (a) simulated data, (b) equal-size binning for $K = 6$, (c) optimal binning obtained by minimising J_{KM} for $K = 6$.

J_{KM} is a least-squares criteria that supposes we are dealing with a homoscedastic model, i.e., the data spread (with respect to time) inside each cluster is similar. This is not always the case, as for example in Fig. 7(a). The combined variability of the first two clusters is similar to that of each of the third, fourth and fifth, whereas the variability of the sixth cluster is significantly greater than all the others. In this case, the J_{KM} criteria may not be optimal; Fig 7(b) shows that it groups the first two clusters together, and splits the sixth cluster in two.

To avoid this, let us introduce a model to better take into account heteroscedacity. First, remark that (9) can be rewritten $J_{\mathrm{KM}}(I) = \sum_{k=1}^{K} n_k \sigma_k^2$, where σ_k^2 is the empirical variance of the points in I_k, i.e.,

$$\sigma_k^2 = \frac{1}{n_k} \sum_{i \in I_k} (x_i - \overline{x}_k)^2 .$$

We propose to modify J_{KM} by introducing a further parameter:

$$\widetilde{J}_{\mathrm{KM}}(I,\beta) = \sum_{k=1}^{K} n_k \left(\sigma_k^2\right)^{\beta}. \tag{10}$$

Remark 1. When $\beta = 1$, we have $\widetilde{J}_{\mathrm{KM}}(I,1) = J_{\mathrm{KM}}(I)$.

Remark 2. If we develop (10) for $0 < \beta < 1$, we obtain:

$$\begin{aligned}
\widetilde{J}_{\mathrm{KM}}(I,\beta) &= \sum_{k=1}^{K} n_k e^{\beta \log(\sigma_k^2)} \\
&= \sum_{k=1}^{K} n_k (1 + \beta \log(\sigma_k^2) + o(\beta)) \\
&= n + \beta \sum_{k=1}^{K} n_k \log(\sigma_k^2) + o(\beta).
\end{aligned}$$

For a fixed data set $\{x_1, \ldots, x_n\}$, there therefore exists some small $\beta_0 > 0$ such that for $\beta \leq \beta_0$, minimising $\widetilde{J}_{\mathrm{KM}}(I,\beta)$ is equivalent to minimising

$$\sum_{k=1}^{K} n_k \log(\sigma_k^2).$$

This is no other than a criteria typically used to perform segmentation of signals with respect to change in *variance* [4]. Thus, moving β from β_0 to 1 gives a spectrum of solutions that pass from favoring changes in variance to changes in mean. Fig. 7(b) shows the binning obtained by minimising $\widetilde{J}_{\mathrm{KM}}(I,1)$. Then, as β is set closer and closer to 0, more emphasis is made on selecting bins with differing variability. Fig. 7(c) shows an intuitively optimal binning, obtained by minimising $\widetilde{J}_{\mathrm{KM}}(I,0.2)$. Exactly the same binning is obtained with any value of β in [0.05 , 0.35].

3.5 Selection of the number of bins

Of course, for any given number of bins K, such a binning can be calculated. The question then arises as to which K to choose. We propose here

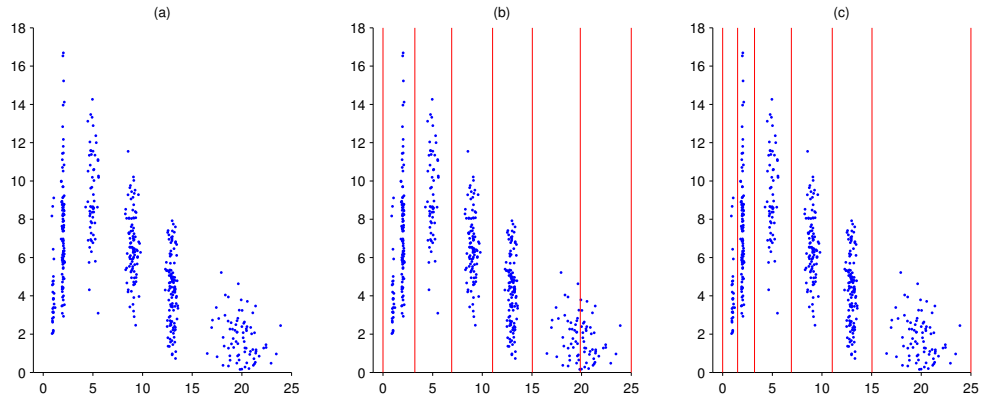

Figure 7: (a) simulated data, (b) binning minimising $\widetilde{J}_{\mathrm{KM}}(I, 1)$, (c) binning minimising $\widetilde{J}_{\mathrm{KM}}(I, 0.2)$.

to automatically select the number of bins using a model selection approach with the following penalised criteria:

$$U(I, \beta, \lambda) = \log\left(\widetilde{J}_{\mathrm{KM}}(I, \beta)\right) + \lambda\beta K(I), \tag{11}$$

where $K(I)$ is the number of bins in binning I. We choose the I (and thus the K) that minimises $U(I, \beta, \lambda)$ for λ fixed. The larger λ is, the fewer the number of bins selected. Extensive numerical trials suggest the use of $\lambda = 0.3$. The β term is included in the penalty because $\log\left(\widetilde{J}_{\mathrm{KM}}(I, \beta)\right)$ decreases, as a first approximation, like a linear function of β.

4 Discussion

In many applications we are confronted with the need to take a set of data and break it up into subsets, and the context tells us what a typical subset should look like. We aim to encode this knowledge as a mathematical criteria that can automate the process of selecting subsets, or more precisely here, perform segmentation of ordered data.

In the first part of the article, we have presented two methods for data-driven reconstruction of unknown piecewise-constant densities using segmentation approaches. The first method attempts to find change-points in the

slope of the inverse of the empirical cumulative distribution function, whereas the second uses a maximum likelihood approach. The set of change-points are then considered to be the predicted bin edges of the density function, allowing an empirical density function to be constructed. Numerical experiments indicate that the maximum likelihood approach has significant computational and performance advantages, and we recommend its use over the change-point detection approach.

In the second part of the article, we considered the general problem of constructing a good histogram representation of a 1-d finite data set for any chosen subject of interest, though with a particular application in mind: Visual Predictive Checks (VPCs) for pharmacometric modelling. Here, real data have to be binned into specific time intervals because otherwise, the confidence intervals we are trying to predict may exhibit overly "bumpy" patterns, making visual interpretation difficult. A good (histogram) binning should aim to be made up of a set of bins which give a reasonable "discrete summary" of the data, and we show that for VPCs, simple automatic binning strategies such as equal-width and equal-size are not appropriate. Consequently, we developed a method - also based on a segmentation approach - that bridges the gap between 1-d K-means and segmentation based on finding change-points in variance.

The two parts of the article are unified by a common mathematical framework whereby a real-valued function g is calculated on all $n(n+1)/2$ possible contiguous subsets of the original data and then the minimum of the sum of these values for all possible partitions I is calculated. The optimal partition induces the set of bin-edges in each case. Using dynamic programming techniques, this minimisation can be done practically in $\mathcal{O}(n^2)$ rather than $\mathcal{O}(n^K)$ time. The method we have presented is thus valid for any practical application where a suitable function g can be defined.

Lastly, we have implemented a model-selection approach whereby the number of true bins K is not necessarily known in advance and must also be predicted. We found that it is in general a hard problem to estimate K, and that for piecewise-constant density reconstruction, a penalty of the form $\text{pen}(K) = K * (\beta_1 + \beta_2 \log(n))$ outperformed BIC, the non-asymptotic approach of Birgé and Massart [25] and various other adaptive procedures for model selection for piecewise-constant densities. Based on extensive numerical trials, we were also able to propose a well-calibrated penalised criteria to select the number of bins for the VPC application.

References

[1] A.K. Jain, M.N. Murty and P.J. Flynn. Data Clustering: A Review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[2] A. Sanders. Real-time discrimination of broadcast speech/music. *Proc. ICASSP-96*, 2:993–996, 1996.

[3] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. *Proc. ICASSP-97*, 2:1331–1334, 1997.

[4] M. Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501–1510, 2005.

[5] A.B. Olshen, E.S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.

[6] P. Hupé, N. Stransky, J.P. Thiery, F. Radvanyi, and E. Barillot. Array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20:3413–3422, 2004.

[7] F. Picard, S. Robin, M. Lavielle, C. Vaisse and J.-J. Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6:27, 2005.

[8] J. Fridlyand, A. Snijders, D. Pinkel, D. Albertson and A. Jain. Hidden markov models approach to the analysis of array CGH data. *J. Multivariate Anal.*, 90:132–153, 2004.

[9] F. Picard, S. Robin, E. Lebarbier and J.-J. Daudin. A segmentation-clustering problem for the analysis of array CGH data. *Biometrics*, 63:758–766, 2007.

[10] J. Huang, A. Gusnanto, K. O'Sullivan, J. Staaf, A. Borg and Y. Pawitan. Robust smooth segmentation approach for array CGH data analysis. *Bioinformatics*, 23(18):2463–2469, 2007.

[11] R. Tibshirani and P. Wang. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9(1):18–29, 2008.

[12] S.M. Kay. *Fundamentals of Statistical Signal Processing*, volume 2. Prentice-Hall, 1998.

[13] A. Hooker, M.O. Karlsson, and E.N. Jonsson. Visual Predictive Check (VPC) using XPOSE. *http://xpose.sourceforge.net/generic_chm/xpose.VPC.html.*

[14] N. Holford. VPC, the visual predictive check – superiority to standard diagnostic (Rorschach) plots. In *PAGE 2005 (http://www.page-meeting.org/?abstract=738)*, 2005.

[15] M.O. Karlsson and R. Savic. Diagnosing model diagnostics. *Clinical Pharmacology & Therapeutics*, 82:17–20, 2007.

[16] M.O. Karlsson and N. Holford. A tutorial on Visual Predictive Checks. In *PAGE 2008 (http://www.page-meeting.org/pdf_assets/8694-Karlsson_Holford_VPC_Tutorial_hires.pdf)*, 2008.

[17] Y. Yano, S.L. Beal, and L.B. Sheiner. Evaluating pharmacokinetic/pharmacodynamic models using the Posterior Predictive Check. *J Pharmacokin Pharmacodynam*, 28(2):171–192, 2001.

[18] N. Molinari, J.-F. Durand and R. Sabatier. Bounded optimal knots for regression splines. *Comput. Stat. Data Anal.*, 45:159–178, 2004.

[19] M. Lavielle. Detection of multiple changes in a sequence of dependent variables. *Stoch. Proc. Appl.*, 83(1):79–102, 1999.

[20] T.M. Post, J.I. Freijer, B.A. Ploeger, and M. Danhof. Extensions to the Visual Predictive Check to facilitate model performance evaluation. *J Pharmacokinet Pharmacodyn*, 35:185–02, 2008.

[21] E. Comets, K. Brendel, and F. Mentré. Model evaluation in nonlinear mixed effect models, with applications to pharmacokinetics. *Journal de la SFdS*, 151:106–128, 2010.

[22] M. Bergstrand, A.C. Hooker, J.E. Wallin, and M.O. Karlsson. Prediction-corrected visual predictive checks for diagnosing nonlinear mixed-effects models. *AAPS J.*, 13(2):143–151, 2011.

[23] D. Wang and S. Zhang. Standardized visual predictive check versus visual predictive check for model evaluation. *J. Clin. Pharmacol.*, 2011.

[24] J.B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press, 1967.

[25] L. Birgé, and P. Massart. Gaussian model selection In *J. Eur. Math. Soc.*, vol. 3, no. 3, pp. 203–268, 2001.