algorithms

MDPI

Article

Directed Clustering of Multivariate Data Based on Linear or Quadratic Latent Variable Models

Yingjuan Zhang [1,†] and Jochen Einbeck [1,2,*,†]

[1] Department of Mathematical Sciences, Durham University, Durham DH1 3LE, UK; yingjuan.zhang@durham.ac.uk

[2] Durham Research Methods Centre, Durham University, Durham DH1 3LE, UK

[*] Correspondence: jochen.einbeck@durham.ac.uk

[†] These authors contributed equally to this work.

Abstract: We consider situations in which the clustering of some multivariate data is desired, which establishes an ordering of the clusters with respect to an underlying latent variable. As our motivating example for a situation where such a technique is desirable, we consider scatterplots of traffic flow and speed, where a pattern of consecutive clusters can be thought to be linked by a latent variable, which is interpretable as traffic density. We focus on latent structures of linear or quadratic shapes, and present an estimation methodology based on expectation–maximization, which estimates both the latent subspace and the clusters along it. The directed clustering approach is summarized in two algorithms and applied to the traffic example outlined. Connections to related methodology, including principal curves, are briefly drawn.

Keywords: clustering; mixture model; latent variable model; dimension reduction; expectation–maximization algorithm; model selection; fundamental diagram

check for updates

Citation: Zhang, Y.; Einbeck, J. Directed Clustering of Multivariate Data Based on Linear or Quadratic Latent Variable Models. *Algorithms* **2024**, *17*, 358. https://doi.org/10.3390/a17080358

Academic Editors: Mario Rosario Guarracino, Laura Antonelli and Pietro Hiram Guzzi

Received: 8 July 2024
Revised: 9 August 2024
Accepted: 13 August 2024
Published: 16 August 2024

1. Introduction

Multivariate clustering methods are now well-developed. A wide range of methods and techniques is available, including empirical methods such as K-means [1,2], parametric methods based on mixture models [3,4], and nonparametric methods based on density modes [5,6]. Questions regarding the choice of model complexity—such as the number of clusters, cluster shape, and bandwidth—are generally well-understood, with statistical software providing efficient implementations to automatically select these parameters. For instance, the R package **mclust** for mixture-based clustering uses the BIC criterion in order to automatically and simultaneously determine the number of Gaussian mixture components and the shapes of these components, selected out of 14 possible variance parameterizations, and allowing for differing degrees of flexibility in terms of volume, shape, and orientation [7]. For a relatively recent (but still valid) overview of contemporary clustering techniques, see [8]. Nonetheless, research on multivariate clustering continues at high intensity: A Google Scholar search on articles since 2015 featuring the terms 'clustering', 'multivariate', and 'methodology' gives, at the time of writing, 19,100 hits. Areas of research activity on clustering in the last 10 years include the development of methods for high-dimensional data, incorporating aspects of variable selection and sparsity [9,10], methods tailored to functional data [11], variations in clustering methods for spatial statistics, including spatially constrained clustering [12], algorithmic variants that allow for faster computation [13,14], and methods that link to soft computing techniques [15], to name a few.

Clustering is generally considered as a process without a sense of direction. In other words, the set of detected clusters is an entirely unordered one. That is, in existing clustering procedures, there is no sense of 'ordering' that can be sensibly assigned to the clusters (apart from, perhaps, by probability mass). In other words, there is usually no way of putting the clusters in relation to each other based on their location within the data space. This is a possible drawback, especially in situations where the relative positioning of the cluster centers (and, hence, of the clusters) is meaningful as it can be considered as being driven by an external variable, which may or may not be observed. Specifically, in many situations, one may argue that the cluster centers are just particular realizations of a latent variable spanning the data space, with the observations essentially constituting multivariate 'noise' around these centers. Such scenarios are particularly prone to occur when the relationship between the variables in question is governed by some physical equation or property. An example of such a situation is the measurement of speed and flow on highways, sometimes referred to as the 'fundamental diagram', where the underlying latent variable can be thought to represent the traffic density. Although the next section is dedicated to detailing this particular application, the reader may wish to glance at Figure 1 at this point to gain some intuition. When clustering this data, the order of the clusters along the 'curved direction' taken by the data cloud is clearly meaningful: there is much more information in the data than would be captured by an unordered list of cluster centers.

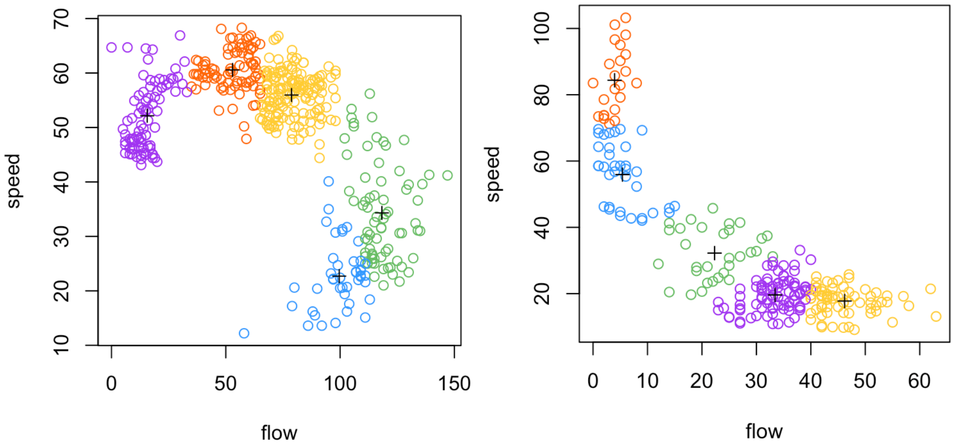

Figure 1. Speed-flow data on a California freeway with K-means clustering. The cluster centers are marked with + symbols. **Left**: Traditional shape of the 'Fundamental diagram', recorded from 9 July 2007 9:00 to 10 July 2007 21:59 by VDS detector 1202263 on California Freeway SR57-N. **Right**: An unusual pattern involving traffic by heavy vehicles on a slow lane, recorded on 9 July 2007 from 0:00 to 19:59 by VDS detector 1213624 on freeway SR57-S. Each point in the plots corresponds to the number of vehicles and average speed over 5-min intervals.

To define the terms more formally, assume the given data $x_i \in \mathbb{R}^m$, $i = 1, \ldots, n$. We consider *clustering* as an algorithmic process that takes the data $\{x_i\}_{1 \leq i \leq n}$ as input, and produces a set of cluster centers, say $c_1, \ldots, c_K \in \mathbb{R}^m$, with *cluster labels* $1, 2, \ldots, K$ as output, where observations x_i are then assigned according to a clustering rule $\hat{k}(x_i)$ with values in $1, \ldots, K$. It is important to realize that, in conventional clustering, it is generally *not* meaningful to establish an ordering relation on the cluster labels; that is, one cannot say that cluster 2 is 'larger' than cluster 1 in any meaningful way. The reason is that any such ordering would be based on the relative positioning of the centers c_k in the m-variate space, but these will not allow for obvious ordering unless $m = 1$. Since this paper is concerned with multivariate clustering, we are only interested in the case where $m > 1$.

We define a *directed clustering* algorithm as a clustering process in which the ordinal relationship of the resulting cluster labels is well-defined; that is, one can say that cluster $k + 1$ is larger (or smaller) than cluster k in a meaningful way. In order to include directional information in the clustering process, one needs to connect the clustering problem with the concept of a latent variable. The idea is to assume the existence of such a latent variable, which parameterizes a curve through m-variate space, passing through $c_1, c_2, ..., c_K$. Since the latent variable is one-dimensional, it enables an ordering relation along it, which can then be inherited by the cluster centers 'sitting' on it and, hence, to their corresponding cluster labels. A suitable latent variable model that can used to achieve this purpose was recently proposed by [16], who suggested approximating highly correlated multivariate data through the use of a one-dimensional latent space, i.e., a straight line, which is parameterized by a single random effect. An estimation of this model was carried out through the nonparametric maximum likelihood approach [17], which is based on a mixture model approximation that facilitates the clustering step. This model achieves clustering of the data and dimension reduction simultaneously. The work by [16] did not introduce the notion of 'directed' clustering, so one novel contribution of the present manuscript is to introduce and implement this concept. However, importantly, it is clear that most data structures are not exactly linear (such as the ones in Figure 1); hence, as a second novel contribution, in this manuscript, we also extend the approach in Ref. [16] to quadratic latent variable models. We propose an algorithm to simultaneously select the number of clusters and choose between a linear and a quadratic latent variable model, and a second algorithm to carry out the actual directed clustering based on this model choice.

Apart from the mentioned example from traffic engineering, potential applications of this method include situations where 'rankings' or 'league tables' are constructed from multivariate data. Instances of such problems are school-effectiveness studies [18] or large-scale international skill surveys [16], where one is not only interested in identifying clusters of similarly performing schools or countries but also in their relative performance to each other, as this is of relevance for resource allocation and related policy decisions. Similar arguments can be made for summaries of economic indicators such as export/import activities or price indices [19], where one may be interested in identifying clusters of the best- or worst-performing countries, in certain senses. Ample areas of potential application can be found in the sciences where relationships are driven by natural laws and processes; for instance, in single-cell RNA sequencing, a common task is the 'cell ordering problem' [20,21], where one attempts to identify clusters of gene co-expressions with the help of a latent variable referred to as 'pseudotime'; in the case of reference [20], this is achieved through a combination of the principal curve [22] and K-means methodology.

This paper proceeds as follows. In Section 2, we introduce the motivating traffic data application in more detail; explicitly making the case for the usefulness of the proposed methodology in this field. In Section 3, we recall the linear latent variable model and then develop the quadratic one, including the estimation approach. Simulation studies that demonstrate the efficiency of the estimation methodology are presented in Section 4. Equipped with the methodology, we can formulate the required algorithms in Section 5. In Section 6, we return to the speed-flow data, producing the directed clustering results for the considered datasets. In Section 7, we draw some parallels between the presented approach and principal curves before the paper is concluded with a Discussion in Section 8.

2. Motivating Application: Speed-Flow Data

Although most drivers typically do not give much thought to the physics of the traffic they are in, they instinctively understand the mechanisms that determine road conditions. Drivers are aware that, when traffic density [vehicles/distance] is low, their speed is essentially only restrained by the speed limit. Once traffic is denser, this does not immediately impact their speed but will increase the flow [or throughput; vehicles/time] of vehicles passing each fixed location. However, when traffic is too dense, the speed is impacted, up to a point when vehicles are jammed and both speed and flow break down synchronously.

This is precisely the story told by the scatterplot depicted on the left side of Figure 1. This plot depicts measurements of traffic speed and flow on the California freeway SR57-N, recorded from 9 July 2007, 9 am, to 10 July 2007, 10 pm, on Lane 5, with 444 observations. The data, which are available as part of a dataset `calspeedflow` in the R package **LPCM** [23], were collected by a loop-detector at a fixed 'vehicle detector station'. Each point in the scatterplot represents the number of vehicles passing over the loop detector within a 5-min interval (flow; x-axis) and the average speed (in miles per hour) of those vehicles in that interval (y-axis).

Direct statistical analyses of speed-flow data are not very widespread in the literature due to potential bimodalities in both coordinate directions, which precludes regression analysis in either orientation. Ref. [24] fitted principal curves to speed-flow data patterns from the California PeMS database. Using data from the same source, but additionally involving a third variable (occupancy, the length of time a vehicle takes to drive over a loop), Ref. [25] demonstrated through the use of an agglomerative clustering algorithm that five clusters can typically be identified. Ref. [26] used K-means clustering to identify the transition point from uncongested to congested conditions in speed-flow data from the metropolitan area of São Paulo, Brazil.

The left panel in Figure 1 shows the clustering of the `calspeedflow` dataset through K-means with five clusters (as in [25]) using the `kmeans` function available in R. While the clusters appear nicely lined up along the data cloud, it is important to note that they are not *actually* ordered—neither K-means nor any other commonly used clustering algorithms inherently provide a method for arranging the clusters in any ordered relationship with each other. In this paper, we follow the approaches of both [24,26], and analyze such data directly in speed-flow space, combining the notions of curvilinear approximation and clustering.

Indeed, considering Figure 1 (left), it is intuitive to postulate the presence of a latent variable, starting in the top left corner of the data and proceeding, continuously and smoothly, to the bottom right of the data cloud. An obvious question is then, why would anyone be interested in such a latent variable? To answer this question, we need to appreciate that the observed pattern is not coincidental: beyond the intuitive arguments given in the first paragraph of this section, one can resort to the fundamental identity of traffic flow, which states that, under certain conditions on the stationarity and homogeneity of the traffic, speed v and flow q are related through the fundamental identity $q = dv$, where d is the traffic density.

As illustrated by [24], traffic density is a monotone transformation of the described latent variable. A mathematical reason for this monotonicity involves the following: Note that the fundamental identity of traffic flow implies $v/q = 1/d$, i.e., the traffic density is determined by the (inverse of) the ratio of speed and flow. Hence, the traffic density at each point is uniquely determined by the slope of the line connecting that point to the origin. For there to be a break in the monotony of traffic density along the latent variable describing the curved data shape, a line through the origin would need to intersect that curve twice. When considering Figure 1 (left), it looks unlikely that this property is strongly violated here, and as we will see in Section 5, it is certainly not violated for the sequence of ordered (directed) cluster centers identified by our methodology.

The data in Figure 1 (right) show a pattern that differs from the typically reported behavior. This dataset was collected from 0:00 to 19:59 on the southbound California Freeway SR57-S, with 240 observations in total, as presented in [24]. Just like the previous dataset, it was originally extracted from the PeMS 7.3 database. As explained in [24], the unusual pattern is likely explained by the presence of heavy vehicles (vehicles with a larger-than-standard gross vehicle weight rating, such as trucks or buses) on a slow lane, where increased speed is counterproductive as it leads to larger stopping distances and, hence, reduces flow. Imagine a latent variable tracing a curvilinear path from the top left to the bottom right of this data cloud. By reasoning similar to the previous example, the assumption that traffic density increases monotonically with respect to the latent variable is plausible. The right panel in Figure 1 also gives the K-means clustering for this dataset.

To summarize, achieving a directed clustering of the data with respect to the latent variable simultaneously results in a directed clustering with respect to traffic density, which is immediately and intrinsically meaningful and interpretable. According to [25], density is “the primary factor to determine the level of service on a freeway”. So, the algorithmic determination of cluster membership ranked by traffic density enables an automatic assessment of the road's operating condition (congested, free-flow, etc.) with immediate relevance for traffic control strategies and intelligent transportation systems. On the contrary, a plain establishment of cluster membership, such as through K-means as in Figure 1, does not allow for identifying the operating condition on the road without additional information about the meaning of these clusters.

Of course, in other applications, it may even be useful to carry out directed clustering along a latent variable that has no such physical meeting. But if it has, as in the present application, the case for such a method is particularly compelling.

We provide the required methodology for directed clustering along curvilinear data structures in the following sections and will return to this example in Section 6.

3. Methodology

3.1. Modeling

Consider multivariate data $x_i \in \mathbb{R}^m$, $m > 1$, $i = 1, 2, \ldots, n$. The model proposed by [16] considers the data generated by a latent variable plus Gaussian noise as follows:

$$x_i = \alpha + \beta z_i + \varepsilon_i, \tag{1}$$

where the univariate random effects z_i are realizations of a random variable Z with density function ϕ (where no distributional assumption is made), $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^m$ are m-variate parameter vectors, and $\varepsilon_i \in \mathbb{R}^m$ are independent Gaussian errors with a diagonal variance matrix $\Sigma = \mathrm{diag}(\sigma_j^2)_{\{1 \le j \le m\}} \in \mathbb{R}^{m \times m}$.

This model is limited by its linear nature, particularly in handling multivariate data structures with the curvature. Given the existence of a non-linear structure in the motivating example data, it becomes apparent that a non-linear model is needed to capture the curvature effectively. In this paper, we propose a quadratic extension of the linear random effect model to allow the fitting of non-linear multivariate data. This extended model can be written as follows:

$$x_i = \alpha + \beta z_i + \eta {z_i}^2 + \varepsilon_i, \tag{2}$$

where $z_i \in \mathbb{R}$ is again a univariate random effect, which can be considered the curve parameterization of a quadratic curve through m-variate space, which is defined by m-variate parameters α, β and η. As before, $\varepsilon_i \sim N_m(0, \Sigma)$ are independent Gaussian errors. For a simplified presentation, we combine both models into a single formulation, as follows:

$$x_i = g(z_i, \theta) + \varepsilon_i, \tag{3}$$

where $g(z, \theta) = \alpha + \beta z$ with $\theta = (\alpha^T, \beta^T)^T$, or $g(z, \theta) = \alpha + \beta z + \eta z^2$ with $\theta = (\alpha^T, \beta^T, \eta^T)^T$ according to Equations (1) or (2), respectively, and proceed (as long as possible) with a unified presentation of the methodology.

Under formulation (3), the conditional probability density function of x_i is a multivariate normal distribution, which can be expressed as follows:

$$f(x_i|z_i, \theta) = \\ (2\pi)^{-m/2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (x_i - g(z_i, \theta))^T \Sigma^{-1} (x_i - g(z_i, \theta)) \right\}.$$

To find the likelihood function, we need to first obtain the marginal probability density function $f(x_i|\theta)$ for observations generated from model (3), which can be written as follows:

$$f(x_i|\theta) = \int f(x_i, z_i|\theta) dz_i = \int f(x_i|z_i, \theta) \phi(z_i) dz_i,$$

where $f(x_i, z_i|\theta)$ is the joint probability distribution of (multivariate, observed) data x_i and (univariate, unobserved) random effects z_i. In principle, the distribution Z of the z_i can be Gaussian or non-Gaussian. If it is non-Gaussian, then one will generally not be able to find an analytical form for the marginal density function. Following the approach for the linear latent variable in [16], we do not make any explicit assumptions regarding the distribution Z of the z_i. Instead, we use the nonparametric maximum likelihood approach for mixture models, intensely discussed by Aitkin et al. [27] and based on theoretical results leading back to [28], in order to deal with the integration over z_i. Specifically, we replace the integral with a set of mass points $z_1, z_2, \dots, z_k$ and their corresponding masses $\pi_1, \pi_2, \dots, \pi_k$, where $k = 1, 2, \dots, K$, so that the approximated marginal probability density function can be written as follows:

$$f(x_i|\theta) \approx \sum_{k=1}^{K} f(x_i|z_k, \theta)\pi_k,$$

which is a mixture of Gaussian components, i.e.,

$$x_i|z_k, \theta \sim N_m(g(z_k, \theta), \Sigma), \tag{4}$$

and mixture probabilities π_k, $k = 1, \dots, K$.

3.2. EM Algorithm

Since the marginal likelihood resulting from specification (4) is not tractable, an EM algorithm is used, which is the most common way of estimating a mixture model [29]. In order to set up the EM algorithm, with component membership serving as the 'missing information', let us define an indicator variable G_{ik}, where it equals 1 if the observation i belongs to mixture component k; otherwise, it equals 0. Defining further $f_{ik} = f(x_i|z_k, \theta)$, it is clear that $P(y_i, G_{ik} = 1) = f_{ik}\pi_k$, and so the probability of the 'complete data' $(x_i, G_{i1}, \dots, G_{iK})$ is as follows:

$$\mathbb{P}(x_i, G_{i1} \dots, G_K) = \prod_{k=1}^{K} (f_{ik}\pi_k)^{G_{ik}}.$$

Hence, the complete log-likelihood is as follows: $\prod_{i=1}^{n} \prod_{k=1}^{K} (f_{ik}\pi_k)^{G_{ik}}$, and the expected complete log-likelihood has the following form:

$$l_c = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}[G_{ik}|x_i] \log(\pi_k f_{ik}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \log \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \log f(x_i|z_k, \theta), \tag{5}$$

where $w_{ik} = \mathbb{E}[G_{ik}|x_i] = \mathbb{P}(G_{ik} = 1|x_i) = \frac{\pi_k f_{ik}}{\sum_l \pi_l f_{il}}$ is the 'posterior' probability of observation i belonging to component k. The expected log-likelihood (5) with the specified expression for $f(x_i|z_k, \theta)$ can be rewritten as follows:

$$\begin{aligned} l_c = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \log(\pi_k) + \sum_{i=1}^{n} \sum_{k=1}^{K} -\frac{1}{2} w_{ik} \log(|\Sigma|) + \sum_{i=1}^{n} \sum_{k=1}^{K} -\frac{m}{2} \log(2\pi) w_{ik} \\ + \sum_{i=1}^{n} \sum_{k=1}^{K} -\frac{1}{2} w_{ik} (x_i - g(z_k, \theta))^T \Sigma^{-1} (x_i - g(z_k, \theta)). \end{aligned} \tag{6}$$

The E-step of the expectation–maximization algorithm is then given by the following:

$$w_{ik} = \frac{\pi_k f_{ik}}{\sum_l \pi_l f_{il}}$$

and the M-step is derived by taking partial derivatives of l_c with respect to each parameter, setting these score equations to 0, and solving them. For the linear latent variable model (1), the resulting estimates, $\hat{\theta} = (\hat{\alpha}^T, \hat{\beta}^T)^T$, $\hat{\Sigma}$, and $\hat{z}_k$, $\hat{\pi}_k$, $k = 1, \dots, K$, are provided in [16], with publicly accessible implementation in the R package [30], and are not repeated here. For the quadratic model (2), these derivations are more involved, and detailed in the next subsection.

3.3. The M-Step for the Quadratic Model

For model (2), taking partial derivatives of (6) w.r.t. α, β, η, and z_k, leads to a complex set of expressions that we provide in Appendix A. However, these expressions involve many matrix inversions and are computationally not very stable to evaluate. Therefore, we follow the approach in [16], where, only within these expressions, we assume that $\hat{\Sigma} \equiv \mathrm{diag}(\sigma^2)$, for some constant σ^2, which does not need to be specified since it cancels out from the resulting simplified update equations for α, β, and η. The resulting simplified estimators are then given by the following:

$$\hat{\alpha} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k - \eta\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^2\right), \tag{7}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \hat{\alpha} - \hat{\eta}\hat{z}_k^2)\hat{z}_k}{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^2}, \tag{8}$$

$$\hat{\eta} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}x_i\hat{z}_k^2 - \hat{\alpha}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^2 - \hat{\beta}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^3}{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^4}. \tag{9}$$

The estimates of z_k will be obtained by solving the cubic equation as follows:

$$\begin{aligned}
(\sum_{i=1}^{n}\sum_{j=1}^{m} 2\cdot w_{ik}\hat{\eta}_j^2){z_k}^3 + 3\cdot(\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ik}\hat{\beta}_j\hat{\eta}_j){z_k}^2 \\
-(\sum_{i=1}^{n}\sum_{j=1}^{m} 2\cdot w_{ik}x_{ij}\hat{\eta}_j - \sum_{i=1}^{n}\sum_{j=1}^{m} 2\cdot w_{ik}\hat{\alpha}_j\hat{\eta}_j - \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ik}\hat{\beta}_j^2)z_k \\
-(\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ik}x_{ij}\hat{\beta}_j - \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ik}\hat{\alpha}_j\hat{\beta}_j) = 0.
\end{aligned} \tag{10}$$

Unfortunately, there is still an added complication since, even when conditional on $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\eta}$, obtaining the analytical form of $\hat{z}_k$ is not possible. However, the score equation (10) is a cubic equation of z_k, for which solvers exist in standard software packages. We use the `polyroot()` function (available in the **base** [31] R package) to solve this cubic equation. It is important to note that the roots of a cubic equation fall into two scenarios: (i) one real root and two complex roots, in which case we consider the real root as the estimate of z_k, and (ii) three real roots, where we always choose the smallest absolute root in the implementation.

The estimates in Equations (7)–(10) still depend on each other, but they do not depend on $\hat{\Sigma}$. Hence, within each M-step of the EM algorithm, one can iterate between them a small number of times, then estimate π_k via $\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} w_{ik}$, and finally estimate the entries of Σ via the following:

$$\hat{\sigma_j}^2 = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_{ij} - \hat{\alpha}_j - \hat{\beta}_j\hat{z}_k - \hat{\eta}\hat{z}_k^2)^2}{n}. \tag{11}$$

The simulation study in the following section demonstrates that this procedure accurately estimates the model parameters. It is also noted that, as in [16], one can work with cluster-dependent variance matrices Σ_k, $k = 1, \ldots, K$, which vary in size and shape. However, in the context of directed clustering, the benefits of doing so are limited. Therefore, we relegate a brief consideration of cluster-dependent variance matrices to Appendix B but continue with the homoscedastic case in the main flow of this paper.

4. Simulation

We conducted a simulation study to evaluate the accuracy of parameter estimation for the quadratic latent variable model (2), based on an implementation of the EM algorithm according to Section 3.3. We consider a simple scenario where x_i are 2-dimensional data generated according to model (4) from $K = 4$ mixture components, with three sample sizes $n = 100$, $n = 300$, and $n = 500$. The true model parameters are displayed in the left column of Table 1. Figure 2 shows the data structures of the data we used for this simulation study. We generate 300 replicates for each sample size, respectively. For each of the 300 replicates (for each sample size), we run the EM algorithm 20 times to select the best estimates with the smallest BIC value; this process aids in selecting a good starting value for the EM algorithm.

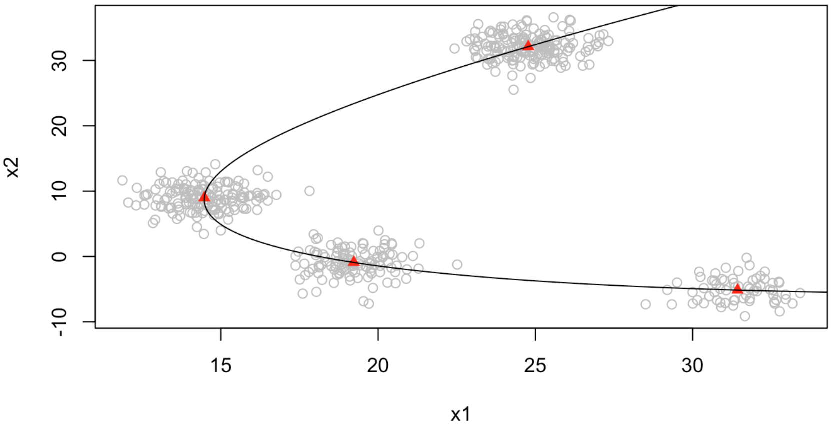

Figure 2. Simulated data with a sample size of 500. The fitted curve is displayed in black, and the mixture centres are given by red triangles.

Table 1. Simulation results for all parameters.

		Average Estimates		
	True	$n = 100$	$n = 300$	$n = 500$
π_1	0.1500	0.1269	0.1269	0.1355
π_2	0.2000	0.1949	0.1954	0.1946
π_3	0.3000	0.2916	0.2963	0.2971
π_4	0.3500	0.3866	0.3814	0.3727
z_1	1.2696	1.2834	1.2847	1.2769
z_2	0.2402	0.2091	0.2021	0.2179
z_3	−0.4461	−0.4635	−0.4579	−0.4497
z_4	−1.0637	−1.0289	−1.0288	−1.0451
α_1	5.0000	5.4230	5.3237	5.3444
α_2	15.0000	14.9822	15.0769	15.0253
β_1	−5.0000	−6.6284	−6.5173	−6.0477
β_2	15.0000	14.1142	14.3244	14.3750
η_1	5.0000	4.8833	4.8010	4.9135
η_2	10.0000	10.4077	10.3825	10.2027
σ_1	1.0000	1.1239	1.1701	1.1103
σ_2	2.0000	2.2411	2.3176	2.2107

The averaged estimates are shown in Table 1, and boxplots of the estimates for some of the parameters are provided in Figures 3–6. Firstly, we observe that as the sample size increases, the boxplots squeeze towards the true values for each parameter, indicating empirically the consistency of the proposed estimation procedure. For the z_k, this is less clearly visible, since even at $n = 100$, the interquartile range (IQR) is already very small.

We also observe that the averaged estimations in Table 1 are generally very close to their true values, except for the β parameters. However, closer inspection of Figure 5 shows that the estimates of β_1 and β_2 are correctly centered; but there is a considerable amount of outliers similarly observed in the boxplots for the η parameters (see Figure 6). We proceed with displaying the mean squared errors (MSEs) for the estimates of the parameters in

Table 2; note that, for β and η, due to the mentioned presence of outliers, we use the trimmed MSE [32] with a trimming parameter of 0.1.

Overall, from these tables and boxplots, we find that the estimators give sensible estimates and that the bias and MSE are reduced with the increase in the sample size.

Table 2. Mean squared errors (MSEs) for all estimated parameters. For β and η, the MSEs are trimmed with the trimming parameter 0.1.

	$n = 100$	$n = 300$	$n = 500$
π_1	0.0033	0.0033	0.0021
π_2	0.0014	0.0009	0.0006
π_3	0.0015	0.0006	0.0005
π_4	0.0060	0.0063	0.0042
z_1	0.0030	0.0035	0.0020
z_2	0.0120	0.0125	0.0071
z_3	0.0106	0.0130	0.0101
z_4	0.0128	0.0146	0.0092
α_1	2.5781	1.6265	1.5941
α_2	0.5989	0.3743	0.3437
β_1	8.0054	6.4488	1.8666
β_2	3.1502	0.0963	0.0190
η_1	0.8834	0.5009	0.0934
η_2	0.3457	0.3020	0.0454
σ_1	0.1428	0.1797	0.1075
σ_2	0.5376	0.6173	0.4181

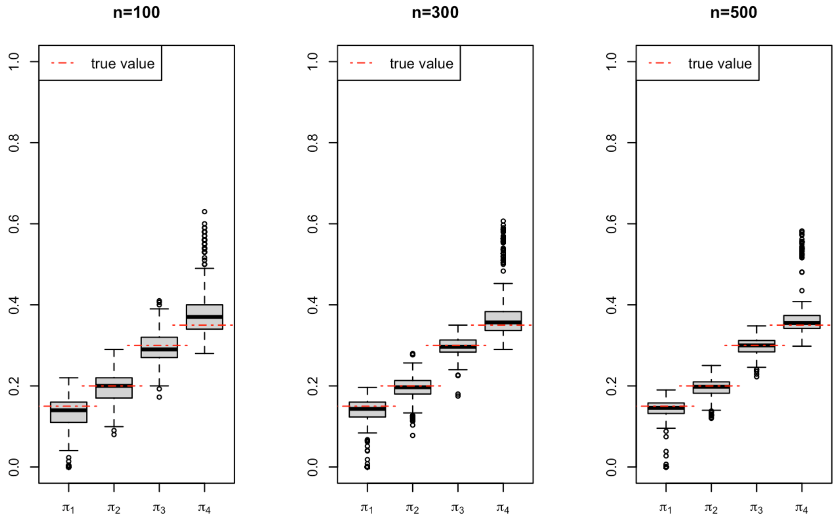

Figure 3. Estimations of parameters $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)^T$ with different sample sizes.

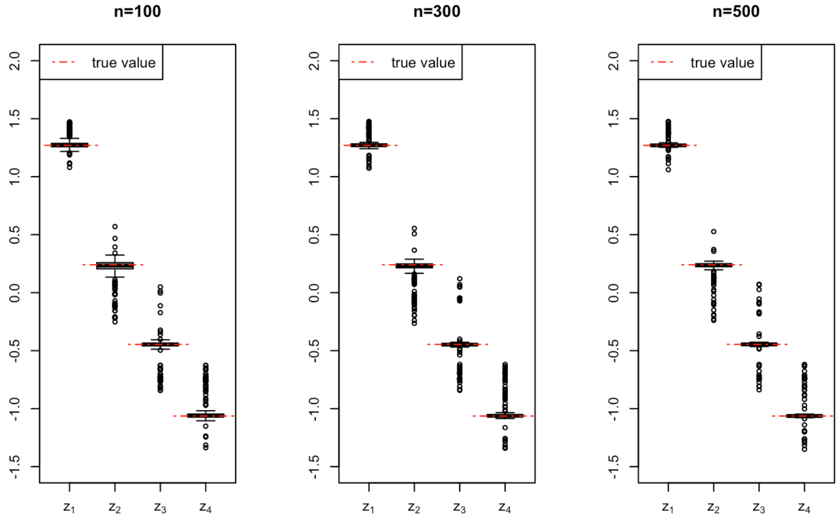

Figure 4. Estimations of parameters $z_k = (z_1, z_2, z_3, z_4)^T$ with different sample sizes.

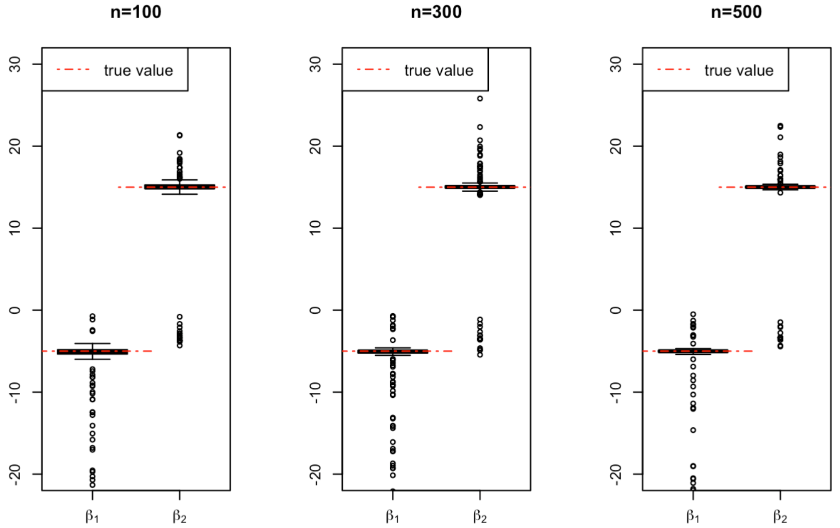

Figure 5. Estimations of parameters $\beta = (\beta_1, \beta_2)^T$ with different sample sizes.

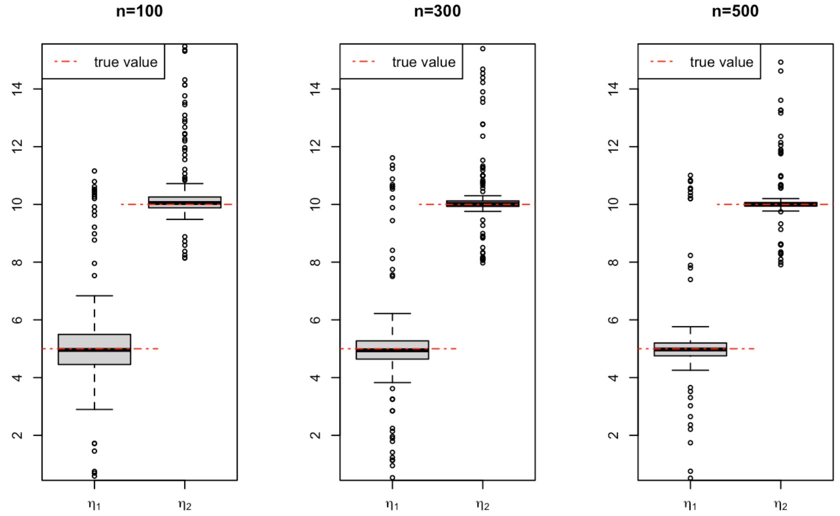

Figure 6. Estimations of parameters $\eta = (\eta_1, \eta_2)^T$ with different sample sizes.

5. Algorithms Required for Directed Clustering

Given a multivariate dataset $x_i \in \mathbb{R}^m$, $m > 1$, $i = 1, 2, \ldots, n$, we first need to decide whether to fit the data with the linear model or the quadratic model. The model selection process is carried out with Algorithm 1. The algorithm requires the specification of a maximum considered number of components, K_{max}. We suggest taking $K_{max} = 10$, as in practical applications there will rarely be a need to use more than 10 components. This is similar to the 'nonparametric maximum likelihood estimation' of standard mixture models; the monograph by Aitkin et al. [27] discusses this approach extensively.

```
Algorithm 1: Model selection
    (i)   Choose an integer $K_{max}$ and fit the linear model (1) and the quadratic model (2)
          separately to the data, for all $K = 2, 3, \ldots, K_{max}$.
    (ii)  For each model type, (1) and (2), select the best-fitting model by identifying the
          value of $K$ yielding minimum BIC values.
    (iii) We compare the BIC values from the best-fitting linear model and the best
          fitting quadratic model. The model with the minimum BIC value will be our
          chosen model.
```

From the best-fitted model according to Algorithm 1, we obtain the posterior probability matrix $W = (w_{ik})_{1 \le i \le n, 1 \le k \le K}$ and the estimated mass points $\hat{z}_1, \ldots, \hat{z}_K$. Each mass point $\hat{z}_k$ has a corresponding mass $\hat{\pi}_k$, and each pair of mass points and its mass has a matching column in the W matrix. The directed clustering of m-dimensional non-linear data $\{x_i\}$ will be obtained according to Algorithm 2.

```
Algorithm 2: Directed clustering
    (i)   Fit the data $\{x_i\}$ according to the best model identified by Algorithm 1.
    (ii)  Order the estimated mass points $\hat{z}_1, \ldots, \hat{z}_K$ in an ascending order.
    (iii) Reorder the columns of $W$ according to the ascending order of mass points.
    (iv)  Perform the clustering rule (e.g., MAP rule, $\hat{k}(x_i) = \max_k w_{ik}$) on the rows of the
          reordered $W$ matrix.
```

Following the strategy above, each observation x_i will be assigned to a cluster k, which is ordered along the curve $g(z, \hat{\theta})$ according to the latent variable, represented by the values of the mass points $\hat{z}_k$.

6. Application on Speed-Flow Data

6.1. Model Selection

We apply Algorithm 1 to the two datasets introduced in Section 4. We fit both the linear model and the quadratic model with different numbers of mixture components K, where $K_{max} = 10$. Tables 3 and 4 show the BIC values for the speed-flow data from the left- and right-hand panels of Figure 1, respectively. Comparing the minimum BIC values of the fitted linear model and the fitted quadratic model indicates that the quadratic model provides a better fit for each of these two datasets; the quadratic model with $K = 4$ is selected for the first dataset, and the quadratic model with $K = 8$ for the second dataset.

Figures 7 and 8 illustrate the fitting process of the quadratic model across the considered range of K values. They show that generally increasing K improves the model fit (decrease bias); however, as is known from the 'usual' nonparametric maximum likelihood estimation [27], once it reaches a certain K value, increasing the number of mass points further will no longer make much improvement to the fitting. From Tables 3 and 4, one can see that the quadratic models achieve their optimal BIC values at larger values of K than the linear models. The reason for this behavior is that the quadratic shape of the curves allows capturing clusters, which would otherwise be masked when simply projecting onto a line; this becomes intuitive when looking at the collection of clusters on the left part of each scatterplot in Figure 8.

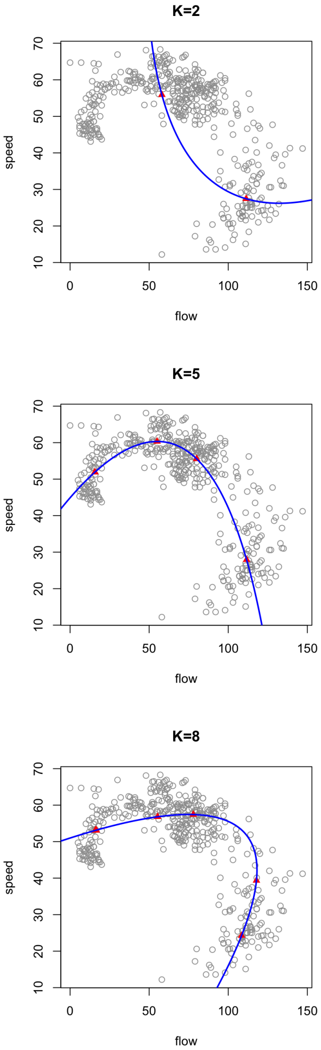
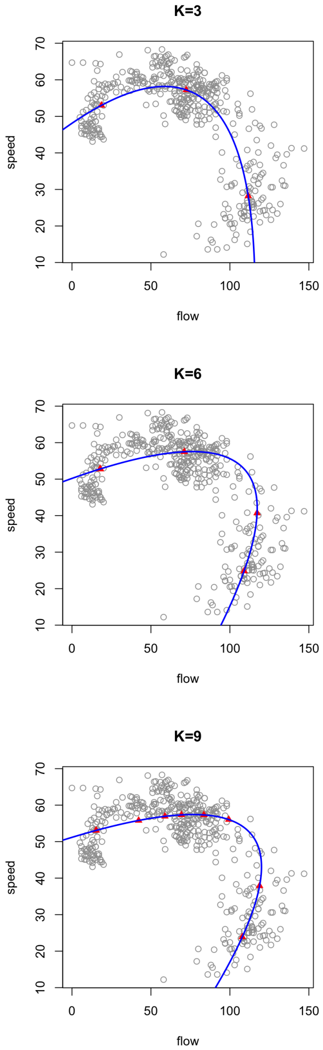
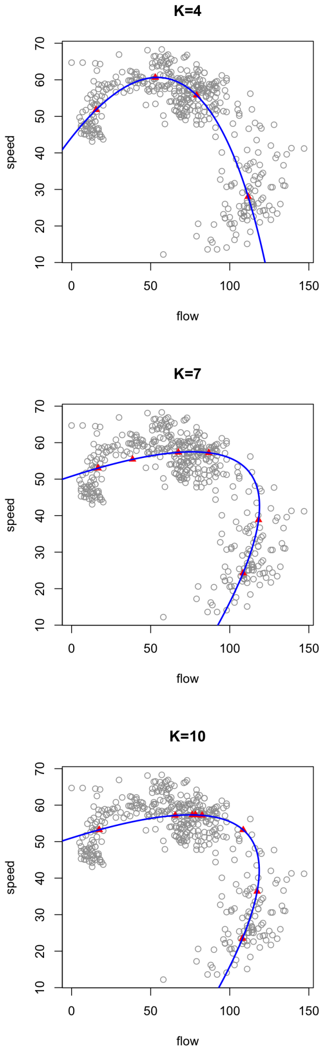

Figure 7. The `calspeedflow` data from the northbound freeway SR57-N, fitted with a quadratic curve (in blue) and mass points represented by red triangles, with different numbers of mixture components.

Table 3. The BIC values for both the linear model and the quadratic model fitted with different numbers of mixture components, for the left-hand data from Figure 1.

	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 8$	$K = 9$	$K = 10$
Linear	**7545.65**	7557.84	7570.03	7582.23	7594.42	7606.61	7618.80	7630.99	7643.18
Quadratic	7557.84	7391.20	**7359.93**	7371.79	7365.24	7364.84	7374.60	7382.78	7401.42

Note: The best fit for each model type is indicated in bold letters.

Table 4. The BIC values for both the linear model and the quadratic model fitted with different mixture components, for the right-hand data from Figure 1.

	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 8$	$K = 9$	$K = 10$
Linear	3831.80	3776.75	**3751.58**	3759.37	3766.83	3781.53	3786.85	3797.42	3808.69
Quadratic	3842.76	3751.23	3700.76	3709.41	3696.52	3694.48	**3689.92**	3696.85	3705.95

Note: The best fit for each model type is indicated in bold letters.

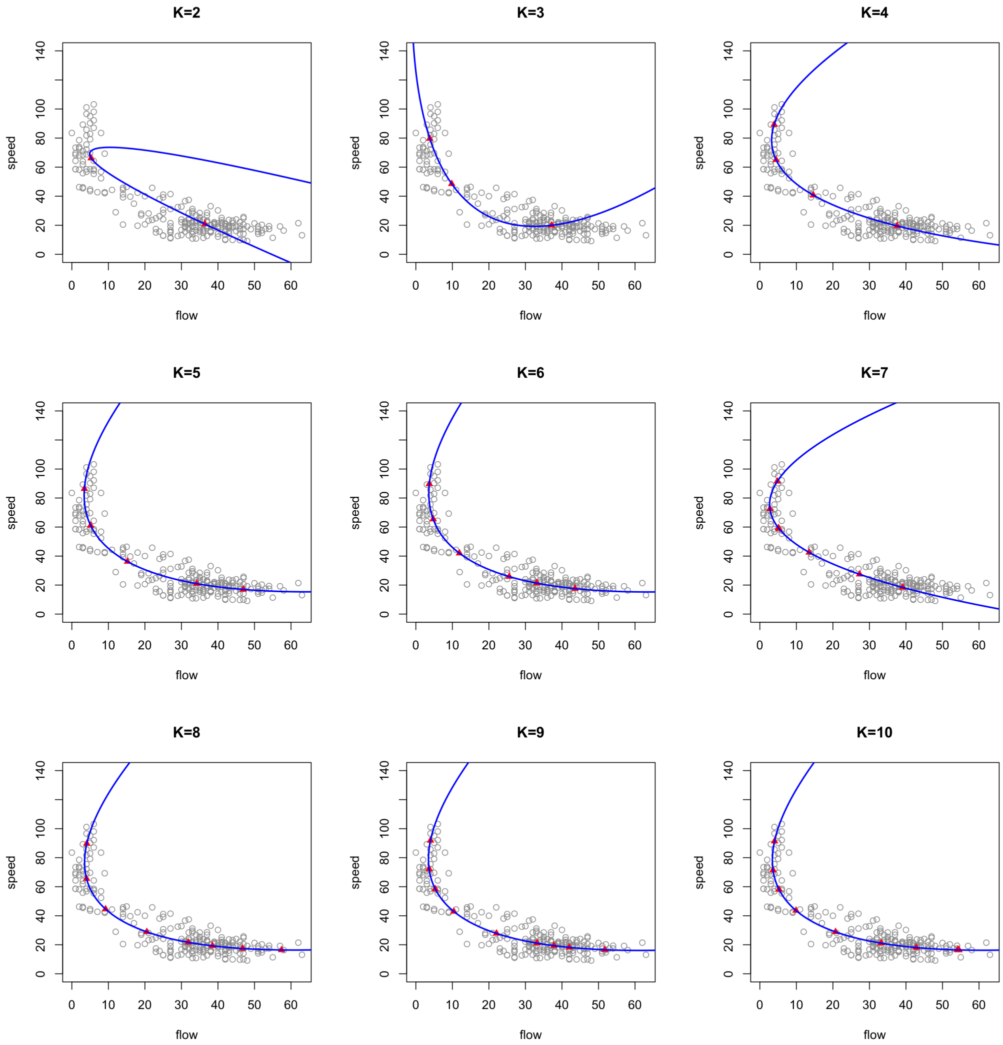

Figure 8. Speed-flow data from the southbound freeway SR57-S, fitted with a quadratic curve (in blue), and mass points represented by red triangles, with different numbers of mixture components.

6.2. Directed Clustering

Now, we demonstrate how Algorithm 2 can be applied to obtain directed clustering of the speed-flow datasets. Based on the output of Algorithm 1, we fit the quadratic model with $K = 4$ and $K = 8$ mass points, respectively, to the two datasets. The estimated mass points $\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_K$ and the posterior probability matrix W are obtained at the convergence of the EM algorithm. These are then reordered according to steps (ii) and (iii) of Algorithm 2. This is followed by applying the maximum a posteriori (MAP) clustering rule according to step (iv). The resulting clusterings are illustrated in the left panels of Figures 9 and 10. One can see how the clusters are lined up along the estimated curve, and ordered with respect to the latent variable that parameterizes the curve.

Returning to the considerations expressed in Section 2, one can relate the clustering to traffic density. This argument can be made explicit by computing the theoretical traffic density at each mass point according to the fundamental diagram, which is given by the following:

$$d_k = \frac{q_k}{v_k} = \frac{g(\hat{z}_k, \hat{\theta})_{[1]}}{g(\hat{z}_k, \hat{\theta})_{[2]}}$$

where the indices [1] and [2] stand for the first and second elements of the vector, respectively. Figure 11 displays the d_k against k. It is visible from this plot that this relationship is indeed monotone, endorsing that the latent variable has a meaningful interpretation in terms of the traffic density; hence, the clusters are ordered from the top left to the bottom right end by increasing traffic density, starting with low densities at the top left end and proceeding toward larger densities in the bottom right end.

For the first dataset, Table 5 reports the ordered cluster labels with the estimated $\hat{z}_k$, as well as the center coordinates $g(\hat{z}_k, \hat{\theta})$ in the data space, and the density values derived as above. It is clear from this table that the clusters represent increasing density (congestion) from $K = 1$ to $K = 4$. Note that the information required to display this table is already available from part (ii) of Algorithm 2.

Additional features and insights can be obtained from the output (iv) of Algorithm 2. Specifically, we can obtain the posterior random effect through the use of ${z_i}^* = \sum_{k=1}^{K} w_{ik}\hat{z}_k$ [33], where w_{ik} represents the posterior probabilities for each observation. The ${z_i}^*$ can serve as one-dimensional summary information for the x_i and can be used for ranking purposes. The right-hand panels in Figures 9 and 10 show the resulting projections, i.e., the segments connecting x_i with $g({z_i}^*, \hat{\theta})$, along the fitted curve for the respective datasets. In practice, it means that each observation (each 5-min interval of speed- and flow values) can be immediately associated with a specific cluster, corresponding to a certain operating condition at the measured location. This information could then be used by intelligent transportation systems to take appropriate action, or perhaps just to flag a 'congestion score' (in the form of the current cluster label) to relevant road management authorities.

Additionally, we can construct a 'league table' of the observations, ranked by the ${z_i}^*$. We illustrate this here only for the first dataset, i.e., the left one in Figure 1, with results shown in Table 6. The table displays the matrix of posterior probabilities, W, with rows reordered according to increasing z_i^*. The values z_i^* are given in the second column of the table. The last column displays the clustering allocation according to the MAP rule. One sees from the table that information provided by the matrix W is much more fine-grained than that provided by the MAP rule. For instance, we see that the observation at the time stamp 03:35:00 is clearly allocated to mass point 1 (a low-density mass point), whereas the one at 05:35:00 is undecided between the mass points with the lowest and second-lowest density.

Due to the presence of 444 time stamps, it is impossible to include all in the table. Therefore, we only show a portion of the full 'league table' in Table 6. Note that the 5-min time periods are not 'ranked' chronologically in the table. The four mass points represent four levels of traffic density, ranging from mass point 1 to mass point 4, corresponding to the time period with the lowest traffic density to the time period with the highest traffic density. Most of the time periods that have been clustered to low traffic density are between 23:00:00 to 05:35:00; most of the time periods between 05:50:00 and 07:15:00 and the time between 19:45:00 and 23:05:00 are assigned to mass point 2. The time periods from the morning rush hour until around 14:00:00 are assigned to the second high traffic density mass point 3, and most of the time between 14:10:00 and 18:35:00 exhibit the highest traffic density.

Table 5. For the dataset from the northbound freeway, the results of the directed clustering algorithm: ordered cluster labels (first column), cluster centers in latent space and data space (second and third columns, respectively), derived density values (final column).

K	$\hat{z}_k$	(q_k, v_k)	d_k
1	−1.14	(15.46, 51.78)	0.30
2	−0.36	(52.95, 60.62)	0.87
3	0.27	(79.06, 55.75)	1.42
4	1.22	(111.62, 28.02)	3.98

Table 6. For the `calspeedflow` data with $K = 4$, posterior probability matrix W with application to classification and ranking. Posterior probabilities: 🟥 $0.10 < p < 0.90$, ▨ $0.90 \leq p < 0.95$, ■ $0.95 \leq p < 1$. (Note that we highlight the largest probability only if it exceeds 0.90.)

Time Stamp	**Posterior Intercept**		**Mass Points**				**MAP**
		k	**1**	**2**	**3**	**4**	
		$\hat{\pi}_k$	**0.18**	**0.19**	**0.39**	**0.24**	
	z_i^*	$\hat{z}_k$	**−1.14**	**−0.36**	**0.27**	**1.22**	
03:35:00	−1.14		1.00	0.00	0.00	0.00	1
03:55:00	−1.14		1.00	0.00	0.00	0.00	1
02:05:00	−1.14		1.00	0.00	0.00	0.00	1
02:35:00	−1.14		1.00	0.00	0.00	0.00	1
⋮	⋮		⋮	⋮	⋮	⋮	⋮
05:35:00	−0.78		0.54	0.45	0.01	0.00	1
23:10:00	−0.63		0.35	0.65	0.00	0.00	2
22:30:00	−0.56		0.27	0.72	0.01	0.00	2
⋮	⋮		⋮	⋮	⋮	⋮	⋮
06:00:00	−0.41		0.07	0.92	0.01	0.00	2
06:15:00	−0.40		0.06	0.93	0.01	0.00	2
⋮	⋮		⋮	⋮	⋮	⋮	⋮
20:05:00	−0.35		0.00	0.98	0.02	0.00	2
21:30:00	−0.34		0.02	0.94	0.04	0.00	2
⋮	⋮		⋮	⋮	⋮	⋮	⋮
20:30:00	−0.07		0.00	0.55	0.45	0.00	2
20:10:00	−0.03		0.00	0.49	0.51	0.00	3
⋮	⋮		⋮	⋮	⋮	⋮	⋮
19:30:00	0.22		0.00	0.09	0.91	0.00	3
10:15:00	0.22		0.00	0.09	0.91	0.00	3
13:05:00	0.22		0.00	0.09	0.91	0.00	3
10:05:00	0.22		0.00	0.09	0.91	0.00	3
⋮	⋮		⋮	⋮	⋮	⋮	⋮
11:30:00	0.26		0.00	0.02	0.98	0.00	3
09:55:00	0.26		0.00	0.01	0.99	0.00	3
13:40:00	0.26		0.00	0.01	0.99	0.00	3
⋮	⋮		⋮	⋮	⋮	⋮	⋮
14:00:00	0.27		0.00	0.00	1.00	0.00	3
⋮	⋮		⋮	⋮	⋮	⋮	⋮
17:10:00	1.22		0.00	0.00	0.00	1.00	4
18:25:00	1.22		0.00	0.00	0.00	1.00	4
17:40:00	1.22		0.00	0.00	0.00	1.00	4
17:30:00	1.22		0.00	0.00	0.00	1.00	4
17:55:00	1.22		0.00	0.00	0.00	1.00	4
18:00:00	1.22		0.00	0.00	0.00	1.00	4

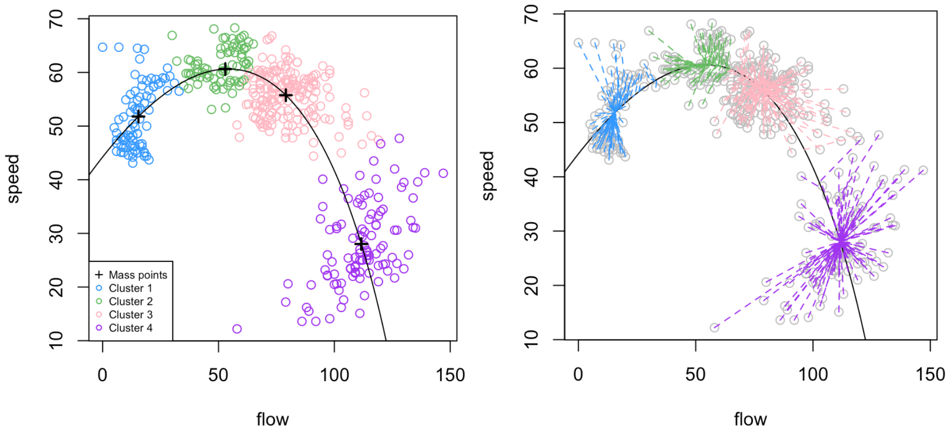

Figure 9. For the `calspeedflow data`, **left panel**: clustering of the speed-flow data using the MAP rule; **right panel**: clustered projections (dashed lines) of the the speed-flow data.

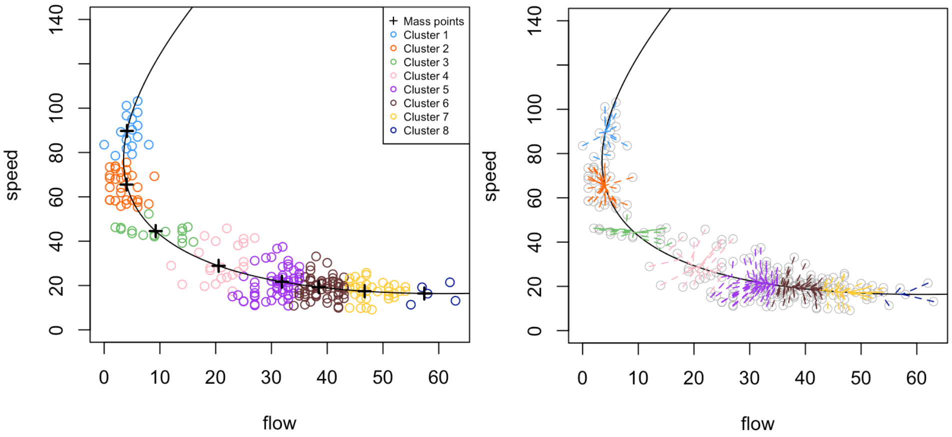

Figure 10. For the second speed-flow dataset, on the southbound freeway SR57-S, **left panel**: clustering of the speed-flow data using the MAP rule; **right panel**: clustered projections (dashed lines) of the speed-flow data.

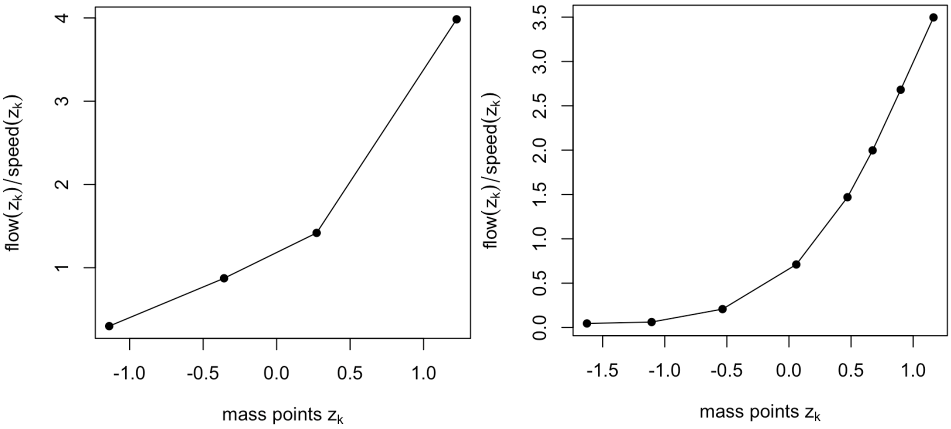

Figure 11. The ratio of flow over speed against z_k for the two speed-flow datasets, presented in the same order as in Figure 1.

7. Comparison with Principal Curves

The fitted curves shown in the previous sections suggest that this methodology could be seen as another approach to estimating principal curves [22]. Both principal curves and the proposed methodology aim to approximate multivariate datasets with smooth curves parameterized by a one-dimensional latent variable.

It is clear that the model formulation (3), which encompasses both the linear and the quadratic case, can be seen as special cases of a generative data model

$$x_i = g(z_i) + \varepsilon_i, \tag{12}$$

where $g : \mathbb{R} \mapsto \mathbb{R}^m$ is a smooth curve. This model was formulated by Hastie and Stuetzle [22] as their base model for their approach to estimating principal curves, informally defined as 'smooth curves passing through the middle of a dataset'. Their estimation approach is based on the self-consistency property, loosely meaning that each point on the curve is the mean of all data points, which project onto that point. Under this approach, g is not actually the principal curve for data generated from the model (12). An example illustrating this point is provided by [34]. However, conceptually, it is still right to think of 'principal curves' being a concept trying to estimate g in (12), and other principal curve definitions, such as by Ref. [34], do not suffer from this conceptual issue. Some principal curve estimation approaches, including local principal curves [35], do not even postulate an underlying model at all. In either case, the similarity between (3) and (12) justifies a brief comparative look at our approach in relation to principal curves.

Here, we restrict to the first of the two speed-flow datasets. We choose to fit the quadratic model with $K = 4$, which leads to a minimum BIC value of 7359.93, for comparison with the principal curves. We also fit the data with a Hastie and Stuetzle principal curve, using the R package **princurve** [36], and a local principal curve (LPC), using R package **LPCM** [23]. The results are shown in Figure 12. We observe that both the fitted HS principal curve (top left panel) and the curve arising from the quadratic latent variable model (bottom panel) fail to fully capture the curvature of the data. In contrast, the local principal curve (top right panel) manages to pull into the endpoint associated with maximum traffic density. One can still argue that the latent variable model produces a sensible result: The methodology aims to locate the mixture centers, which it has done convincingly, and other points around them are considered just noise (belonging to the respective clusters).

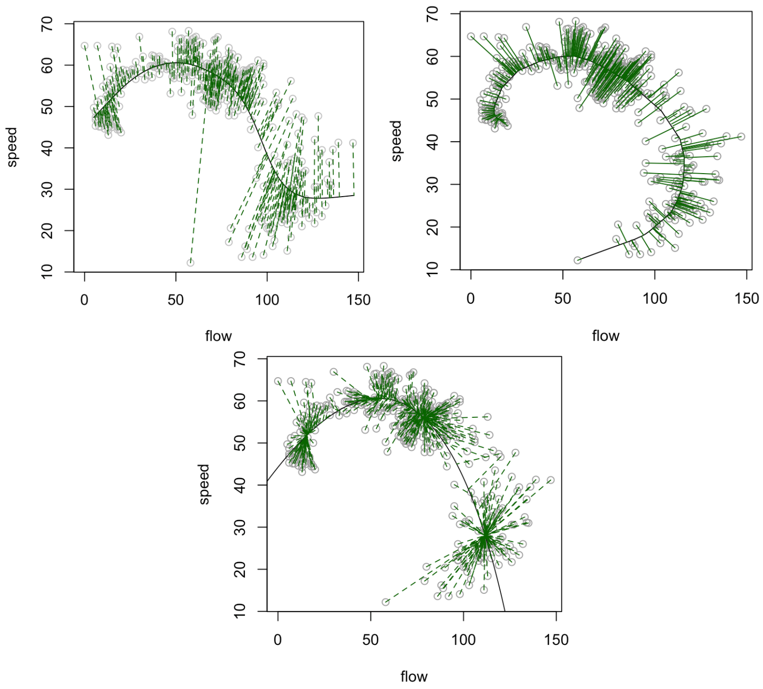

Figure 12. Top: Hastie–Stuetzle (**left**, solid) and local principal curve (**right**, solid) with orthogonal projections (dashed) of the data onto the fitted curve; **bottom**: the fitted curve (solid) resulting from the quadratic latent variable model, with projections (dashed) onto the curve defined by the realization of the latent variable at the posterior random effect of the respective data point.

The main difference between the approaches lies in the way the projections are carried out, as is evident in Figure 12. While under the HS and LPC frameworks, one has orthogonal projections; under the latent variable approach proposed in this paper, one does not, since the dashed lines $x_i - g(z_i^*, \hat{\theta})$ are driven toward the cluster centers.

Apart from qualitatively comparing the fitted curves, it is not fair to use a quantitative assessment, such as the goodness of fit, to compare them, given the different characteristics of the projections from these methods.

8. Discussion

We presented an approach for clustering multivariate data that enables the ordering of clusters along a curve parameterized by a latent variable, estimated alongside the clustering process. We illustrated this concept in the context of linear and quadratic latent variable models. We showed how to choose between these models, how to implement the estimators, and how to interpret the results. Particular attention was devoted to an application in traffic engineering, where the latent variable had a meaningful interpretation, corresponding to the traffic density. The application of the proposed methodology is not restricted to this field; it has potential applicability across various scientific fields. While the idea of a latent variable representing some 'underlying governing process' seems more natural in the physical sciences, the concept of latent variables is home to many other sciences, including fields like education or economics, as alluded to in the introduction. However, it is not actually necessary for the researcher to identify or justify the existence of a governing process in order to use the presented approach. At a minimum, there should be a belief that, grounded in subject matter expertise, the relative ordering of the clusters is informative. In doing so, the existence of a latent variable spanning the centers is postulated—and then the methodology estimates it.

Our work contributes to the development of clustering algorithms that account for the governing process underlying the data at hand. As pointed out by a referee, mathematical clustering models are ineffective if they do not consider important driving factors that contribute to the generation of patterns. The proposed methodology allows the identification of a potentially existing latent variable driving the data-generating process and also allows for the further processing of this latent variable through its posterior random effects. This could include applications such as regression or correlation with actually measured physical variables. Conceptually, it would be even more desirable to directly include the knowledge of the governing process of the data in the clustering algorithm. To a certain extent, this could be achieved by adjusting the clustering process for the presence of influencing variables. For instance, in the context of the traffic examples, one may want to adjust for spatial or temporal factors. This would require the inclusion of covariates into the model (3). In the case of the linear model relationship (1), this was developed in [16], and could be extended to the more general framework presented here.

It is apparent that the restriction to linear and quadratic latent variable models is a possible limitation of the presented methodology. To overcome this restriction, more complex shapes of latent variable models could be considered, such as:

$$x_i = \sum_{j=0}^{p} \alpha_j B_j(z_i) + \varepsilon_i, \tag{13}$$

where $x_i \in \mathbb{R}^m$ and $z_i \in \mathbb{R}$ play the same roles as before, $B_j(z_i)$ can be regarded as any real-valued basis function, such as a B-spline or polynomial, and α_j is a m-variate parameter vector. These still fit into the notational framework defined by (3), but now with $\theta = (\alpha_0^T, \ldots, \alpha_p^T)^T$. The potential power of such an approach becomes evident when considering the LPC fit in Figure 12, which shows that there is still considerable scope for improvement in fitting capability for sufficiently complex data patterns. However, given the difficulties of even estimating the quadratic model (2), requiring a combination of iterative statistical algorithms (EM) and numerical root finders, it will be a considerable challenge to devise an algorithm to estimate this model in full generality.

To prevent doubt, we would like to highlight that we do not advocate this methodology as a superior or favorable clustering technique for generic clustering applications. While the added structure embedded into the model may lend, in some occasions, computational stability to the clustering process, there is no general advantage of using this approach if one is not interested in the directional aspects of the clusters. One reason for this is that the linear or quadratic models constrain the flexibility in positioning the cluster centers, which generally will not lead to more precise or better-fitting clusters than conventional approaches.

Apart from the directional clustering itself, further applications of this approach arise from the presented methodology such as the construction of rankings or 'league tables'. While it could be felt that the construction of the league table for the traffic dataset was a slightly contrived exercise—for other applications, such as when considering sets of educational attainment indicators [16], or economic indicators such as import/export activity, it will not. We leave such considerations for interested researchers to ponder over.

Author Contributions: Conceptualization, Y.Z. and J.E.; methodology, Y.Z. and J.E.; software, Y.Z.; validation, Y.Z. and J.E.; formal analysis, Y.Z.; investigation, Y.Z. and J.E.; resources, Y.Z. and J.E.; data curation, Y.Z. and J.E.; writing—original draft preparation, Y.Z. and J.E.; writing—review and editing, Y.Z. and J.E.; visualization, Y.Z.; supervision, J.E.; project administration, J.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The `calspeedflow` data from freeway SR57-N are available in the R package **LPCM** [23]. The speed-flow data from VDS 1213624 on 9 July 2007, on freeway SR57-S, are part of a dataset provided in Ref. [24].

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

The original expressions for α, β, and η are written as the following, derived by equating partial derivatives of the expected complete log-likelihood (6) to zero:

$$\hat{\alpha} = \left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}^{-1}\right)^{-1}\left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}^{-1}(x_i - \hat{\beta}\hat{z}_k - \hat{\eta}\hat{z}_k^2)\right), \tag{A1}$$

$$\hat{\beta} = \left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}^{-1}\hat{z}_k^2\right)^{-1}\left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}^{-1}(x_i - \hat{\alpha} - \hat{\eta}\hat{z}_k^2)\hat{z}_k\right), \tag{A2}$$

$$\hat{\eta} = \left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}^{-1}(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)({z_k}^2)^T\right)\left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}^{-1}{z_k}^4\right)^{-1}. \tag{A3}$$

The original expression for the cubic equation of z_k has the following form:

$$\begin{aligned}(\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{2\cdot w_{ik}\hat{\eta}_j^2}{\hat{\sigma}_j}){z_k}^3 + 3\cdot(\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{w_{ik}\hat{\beta}_j\hat{\eta}_j}{\hat{\sigma}_j}){z_k}^2\\ -(\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{2\cdot w_{ik}x_{ij}\hat{\eta}_j}{\hat{\sigma}_j} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{2\cdot w_{ik}\hat{\alpha}_j\hat{\eta}_j}{\hat{\sigma}_j} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{w_{ik}\hat{\beta}_j^2}{\hat{\sigma}_j})z_k\\ -(\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{w_{ik}x_{ij}\hat{\beta}_j}{\hat{\sigma}_j} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{w_{ik}\hat{\alpha}_j\hat{\beta}_j}{\hat{\sigma}_j}) = 0.\end{aligned} \tag{A4}$$

Appendix B

For the linear version of model (3), it is possible to use different diagonal variance matrices for different mixture components, as implemented by [16]. In the quadratic version, when we use different diagonal variance matrices for different mixture components, the only change is that Σ will become Σ_k in the log-likelihood (6). The estimators for α, β, η, and the cubic equation for z_k used in the M-step are the same as in Equations (7)–(10). The estimator for the diagonal elements of Σ_k used in the M-step takes the following form:

$$\hat{\sigma}_{jk}^2 = \frac{\sum_{i=1}^n w_{ik}(x_{ij} - \alpha_j - \beta_j z_k - \eta z_k{}^2)^2}{\sum_{i=1}^n w_{ik}}. \tag{A5}$$

When we apply the quadratic model with this type of variance parameterization to the first speed-flow dataset with both $K = 4$ and $K = 5$, we obtain similar clustering results as before, as shown in Figure A1, with $K = 4$ leading to the smallest BIC value. However, we observe that when $K = 5$, it leads to an undesired clustering result where Cluster 5 is surrounded by data points that belong to Cluster 4. This can cause problems for the interpretation of the ordered clustering along the fitted curve.

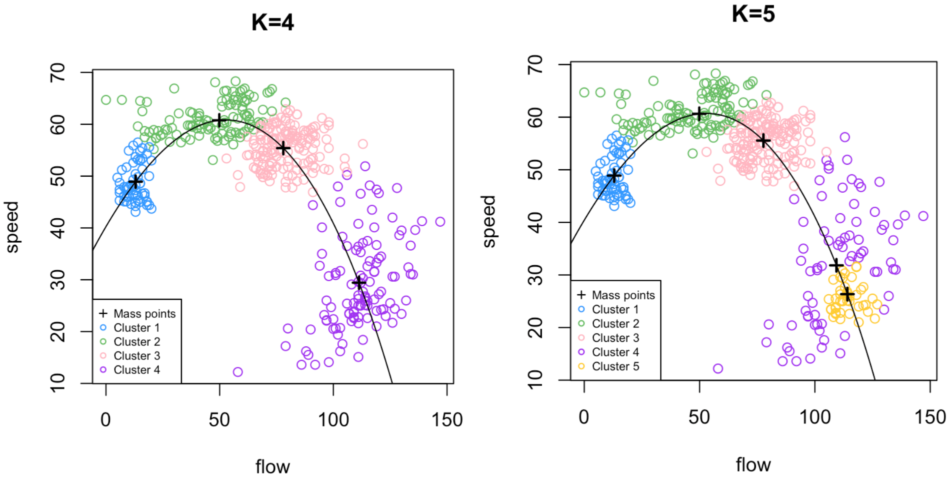

Figure A1. The clustering of `calspeedflow` data using the variance parametrization with different diagonal matrices for different mixture components for $K = 4$ and $K = 5$.

References

1. MacQueen, J.B. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume 1: Statistics*; University of California Press: Berkeley, CA, USA, 1967; pp. 281–297.
2. Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; Abuhaija, B.; Heming, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.* **2023**, *622*, 178–210. [CrossRef]
3. Fraley, C.; Raftery, A.E. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* **2002**, *97*, 611–631. [CrossRef]
4. McNicholas, P.D. Model-based clustering. *J. Classif.* **2016**, *33*, 331–373. [CrossRef]
5. Cheng, Y. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **1995**, *17*, 790–799. [CrossRef]
6. Menardi, G. A review on modal clustering. *Int. Stat. Rev.* **2016** *84*, 413–433. [CrossRef]
7. Scrucca, L.; Fraley, C.; Murphy, T.B.; Raftery, A.E. *Model-Based Clustering, Classification, and Density Estimation Using Mclust in R*; Chapman and Hall/CRC: New York, NY, USA, 2023; ISBN 978-1032234953. Available online: https://mclust-org.github.io/book (accessed on 5 June 2024).
8. Hennig, C.; Meila, M.; Murtagh, F.; Rocci, R. *Handbook of Cluster Analysis*; CRC Press: Boca Raton, MA, USA, 2015.
9. Celeux, G.; Maugis-Rabusseau, C.; Sedki, M. Variable selection in model-based clustering and discriminant analysis with a regularization approach. *Adv. Data Anal. Classif.* **2019**, *13*, 259–278. [CrossRef]
10. Liu, T.; Lu, Y.; Zhu, B.; Zhao, H. Clustering high-dimensional data via feature selection. *Biometrics* **2023**, *79*, 940–950. [CrossRef]
11. Schmutz, A.; Jacques, J.; Bouveyron, C.; Chéze, L.; Martin, P. Clustering multivariate functional data in group-specific functional subspaces. *Comput. Stat.* **2020**, *35*, 1101–1131. [CrossRef]
12. Fouedjio, F. Clustering of multivariate geostatistical data. *WIREs Comput. Stat.* **2020**, *12*, e150. [CrossRef]

13. Deng, C.-H.; Zhao, W.-L. Fast k-Means Based on k-NN Graph. In Proceedings of the IEEE 34th International Conference on Data Engineering (ICDE), Paris, France, 16–19 April 2018; pp. 1220–1223.
14. Zhao, M.; Jha, A.; Liu, Q.; Millis, B.A.; Mahadevan-Jansen, A.; Lu, L.; Landman, B.A.; Tyska, M.J.; Huo, Y. Faster Mean-shift: GPU-accelerated clustering for cosine embedding-based cell segmentation and tracking. *Med. Image Anal.* **2021**, *71*, 102048. [CrossRef]
15. Abdel-Maksoud, E.; Elmogy, M.; Al-Awadi, R. Brain tumor segmentation based on a hybrid clustering technique. *Egypt. Inform. J.* **2015**, *16*, 71–81. [CrossRef]
16. Zhang, Y.; Einbeck, J. A Versatile Model for Clustered and Highly Correlated Multivariate Data. *J. Stat. Theory Pract.* **2024**, *18*, 5. [CrossRef]
17. Aitkin, M. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **1999**, *55*, 117–128. [CrossRef] [PubMed]
18. Aitkin, M.; Longford, N. Statistical modelling issues in school effectiveness studies. *J. R. Stat. Soc. Ser. A (Gen.)* **1986**, *149*, 1–26. [CrossRef]
19. Zayed, M.; Einbeck, J. Constructing Economic Summary Indexes via Principal Curves. In Proceedings of the COMPSTAT 2010, Paris, France, 22–27 August 2010; pp. 1709–1716.
20. Campbell, K.; Ponting, C.P.; Webber, C. Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell RNA-seq profiles. *bioRxiv* **2015**, 027219. [CrossRef]
21. Hou, W.; Ji, Z.; Chen, Z.; Wherry, E.J.; Hicks, S.C.; Ji, H. A statistical framework for differential pseudotime analysis with multiple single-cell RNA-seq samples. *Nat. Commun.* **2023**, *14*, 7286. [CrossRef]
22. Hastie, T.; Stuetzle, W. Principal curves. *J. Am. Stat. Assoc.* **1989**, *84*, 502–516. [CrossRef]
23. Einbeck, J.; Evers, L. LPCM: Local Principal Curve Methods. R Package Version 0.47-4. Available online: https://CRAN.R-project.org/package=LPCM (accessed on 5 March 2024).
24. Einbeck, J.; Dwyer, J. Using principal curves to analyse traffic patterns on freeways. *Transportmetrica* **2011**, *7*, 229–246. [CrossRef]
25. Xia, J.; Chen, M. A nested clustering technique for freeway operating condition classification. *Comput.-Aided Civ. Infrastruct. Eng.* **2007**, *22*, 430–437. [CrossRef]
26. Riente G.U.; Setti, A.J. Speed-Flow Relationship and Capacity for Expressways in Brazil. In *Innovative Applications of the Highway Capacity Manual 2010*; Transportation Research Circular, E-C190; Transportation Research Board: Washington, DC, USA, 2014; p. 10.
27. Aitkin, M.; Francis, B.; Hinde, J.; Darnell, R. *Statistical Modelling in R*; Oxford University Press: Oxford, UK, 2009.
28. Laird, N. Nonparametric Maximum Likelihood Estimation of a Mixing Distribution. *J. Am. Stat. Assoc.* **1978**, *73*, 805–811 . [CrossRef]
29. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **1977**, *39*, 1–22. [CrossRef]
30. Zhang, Y; Einbeck, J. mult.latent.reg: Regression and Clustering in Multivariate Response Scenarios. R Package Version 0.1.7. Available online: https://CRAN.R-project.org/package=mult.latent.reg (accessed on 22 March 2024).
31. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2024. R Package Version 4.4.1. Available online: https://www.R-project.org/ (accessed on 24 June 2024)
32. DiTraglia, F. fmscr. R Package Version 0.1. Available online: https://github.com/fditraglia/fmscr/ (accessed on 24 July 2024).
33. Aitkin, M. Empirical Bayes shrinkage using posterior random effect means from nonparametric maximum likelihood estimation in general random effect models. In Proceedings of the 11th International Workshop on Statistical Modelling, Orvieto, Italy, 15–19 July 1996; pp. 87–94.
34. Tibshirani, R. Principal curves revisited. *Stat. Comput.* **1992**, *2*, 183–190. [CrossRef]
35. Einbeck, J.; Tutz, G.; Evers, L. Local Principal Curves. *Stat. Comput.* **2005**, *15*, 301–313. [CrossRef]
36. Cannoodt, R. Princurve 2.0: Fit a Principal Curve in Arbitrary Dimension (June 2018). Available online: https://zenodo.org/records/3351282 (accessed on 5 March 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.