Dynamic nonparametric clustering of multivariate panel data*

Igor Custodio João[a] Julia Schaumburg[a]

André Lucas[a] Bernd Schwaab[b]

[a]Vrije Universiteit Amsterdam and Tinbergen Institute

[b]European Central Bank, Financial Research

July 29, 2022

Abstract

We introduce a new dynamic clustering method for multivariate panel data characterized by time-variation in cluster locations and shapes, cluster compositions, and possibly the number of clusters. To avoid overly frequent cluster switching (flickering), we extend standard cross-sectional clustering techniques with a penalty that shrinks observations towards the *current* center of their *previous* cluster assignment. This links consecutive cross-sections in the panel together, substantially reduces flickering, and enhances the economic interpretability of the outcome. We choose the shrinkage parameter in a data-driven way and study its misclassification properties theoretically as well as in several challenging simulation settings. The method is illustrated using a multivariate panel of four accounting ratios for 28 large European insurance firms between 2010 and 2020.

Key words: dynamic clustering, shrinkage, cluster membership persistence, silhouette index, insurance industry.
JEL classification: C33, C38, G22.

*Email addresses: i.custodiojoao@vu.nl (Custodio João), a.lucas@vu.nl (Lucas), j.schaumburg@vu.nl (Schaumburg), and bernd.schwaab@ecb.europa.eu (Schwaab). Custodio João and Lucas acknowlegde support from the Dutch National Science Foundation (NWO) under grant 406.18.EB.011. Schaumburg acknowlegdes support from the Dutch National Science Foundation (NWO) under grant VI.VIDI.191.169. Address correspondence to Igor Custodio João, VU Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. The views expressed in this paper are those of the authors and they do not necessarily reflect the views or policies of the European Central Bank.

1 Introduction

We propose a new method to cluster multivariate panel data in a dynamic yet stable and economically meaningful way. Building on established cross-sectional clustering methods, such as e.g. k-means clustering, we provide a straightforward and intuitive algorithm to link consecutive cross-sections over time by introducing persistence in cluster assignments via a penalty parameter. This parameter can be chosen in a data-driven way. The approach results in clusters that can be time-varying in location, dispersion, size/composition, and (possibly) in the number of clusters. As our approach ties the different cross-sections together, changes happen gradually over time and cluster switches become more persistent. Both of these features are important in many economic and financial applications.

Many existing econometric approaches for modeling grouped panel data fail to incorporate dynamics in cluster composition, i.e., potential changes in units' cluster membership over time. In economic applications, however, we often expect several units to switch cluster over the sample period, particularly when the time series dimension is large and/or the sample contains periods of stress. Most of the earlier work focuses on clustering entire time-series, while allowing for different types of unobserved heterogeneity in the panel units. Examples include Lin and Ng (2012), Bonhomme and Manresa (2015), Bonhomme et al. (2022), Cheng et al. (2019) and Patton and Weller (2021), who use variations of k-means to iteratively cluster time series and estimate the structure of a linear or nonlinear regression model. A variety of model-based methods to cluster panel data are surveyed in Frühwirth-Schnatter (2011); see also Frühwirth-Schnatter and Malsiner-Walli (2019) for a finite mixture model approach in which the number of mixtures can be different from the number of clusters in the observed data.

Another line of literature studies clusters in panel data by means of repeated cross-sectional clustering; see for example Oliveira and Gama (2012). This allows for cluster switches, but typically generates clusters that are (too) unstable over time as the obtained structure at one point in time has no bearing on the next cross-section. In addition, it is

often unclear how groups can be tracked over time as cluster labeling is partly arbitrary and therefore cluster identification over different cross-sections is difficult; see Frühwirth-Schnatter (2006). Cluster assignment instabilities are also likely to occur when the panel is treated as one large cross-section, to which a hierarchical clustering algorithm is applied that ignores the time-dimension entirely; see for instance Ayadi et al. (2021).

To accommodate economically meaningful cluster switching, while at the same time avoiding too frequent switching behavior that ceases to be interpretable, we propose a new penalized model-free approach. The approach extends the repeated cross-sectional clustering framework of Oliveira and Gama (2012) by adding time-dependence to the cluster assignments. The context we have in mind is one where units that switch, do so gradually and persistently. For instance, when statistically describing firms' business models, we would not expect a unit to cross from group A to B in one period, only to return back from B to A in the next. We label such erratic moves between clusters as "flickering," a feature that we wish to mitigate, while still allowing for flexible dynamics. Specifically, we do so by shrinking observations towards the new (time t) centroid of their previous (time $t-1$) cluster, before grouping all observations into new clusters at time t. To track the identity of the resulting dynamic clusters, we build on algorithmic ideas that identify clusters by maximizing the overlap in cluster membership over time; see, for instance, Kalnis et al. (2005) and Oliveira and Gama (2010).

The penalty parameter that determines the extent of shrinkage in our approach is set in a data-driven way using a modified version of the silhouette index, which is a widely used cluster validation index introduced by Rousseeuw (1987). We first study the properties of this parameter in terms of mis-classification rates in a stylized setting. This allows us to determine optimal values for the penalty parameter analytically. Next, we investigate the approach in a number of challenging simulation settings that are analytically untractable, and verify the theoretical properties also numerically.

We apply our approach to multivariate panel data of $N = 28$ European insurance companies covering $D = 4$ accounting ratios sampled annually between 2010 and 2020. Our sample is close to the set of companies chosen by the European Insurance and Occu-

pational Pensions Authority (EIOPA) for its 2021 insurance sector stress test; see EIOPA (2021, Annex A). We allocate each insurer to one distinct business model (peer) group at each point in time. To our knowledge, our study is the first to do so for the insurance industry. Reliable up-to-date listings of business model peer groups are useful, for example, for prudential supervision. Insurance supervisors, such as the Federal Insurance Office at the U.S. Treasury, or EIOPA as an important part of the European System of Financial Supervision, routinely need to benchmark insurers' capital positions, cost-to-income ratios, and profitability measures. They do so by comparing each firm's incoming data to that of approximately similar other firms; see e.g. SSM (2016) and Lucas et al. (2019) for a discussion in a banking context.

We recover four clusters: re-insurers, life insurers, non-life insurers, and financial conglomerates. The shrinkage parameter is chosen to decrease the number of incidental switches (flickering) while retaining a high overall fit to the data (in terms of silhouette index). Our clustering approach leads to stable cluster allocations over time. By contrast, we verify that the clustering outcomes are visibly more volatile and much harder to interpret economically if no shrinkage is imposed. The results are qualitatively similar whether or not we allow the number of clusters to also vary over time.

Before proceeding, we also mention three other links to earlier literature. First, our work also relates to the literature on segmenting audio recordings; see, for instance, Fox et al. (2011). A typical finding in this literature is that hidden Markov models can produce over-segmentation, that is, too frequent jumping between states, or "flickering." The problem is typically addressed in a Bayesian way by introducing a parameter for self-transitioning and imposing a prior on it. Our approach is different in that we reduce the dynamic problem to a collection of static ones, and introduce a stickiness (or self-transitioning) hyper-parameter chosen by well-known cluster validation criteria. In addition, our approach is model-free and does not require the choice of a prior as in a Bayesian setting.

Second, our work is also related to Catania (2021) and Custodio João et al. (2022). Both papers use a dynamic mixture modeling approach that allows for changes in cluster

membership. The former does so in a score-driven way, and the latter uses a Hidden Markov Model (HMM). Both papers are potentially subject to an over-segmentation or flickering problem; see Fox et al. (2011). Custodio João et al. (2022) address this by enlarging the HMM dynamics with inactive states, ruling out further transitions for some time after an initial transition. Our methodology differs in at least two ways. First, we adopt a more standard, non-parametric approach to the clustering problem without leaning on explicit distributional assumptions as in Catania (2021) and Custodio João et al. (2022). This allows for an easy generalization of our approach to different clustering algorithms. Second, our penalty parameter determining the stickiness in cluster membership is chosen in a data-driven way, whereas the one in Custodio João et al. (2022) is set exogenously using economic arguments.

A final strand of recent literature that is somewhat related to us focusses on grouped heterogeneity in panels and structural breaks; see, for instance, Lumsdaine et al. (2022), Smith (2022) and Wang and Tsay (2019). Even though in these approaches the number and timing of the structural breaks is unknown and can be estimated, a main assumption is that there is a small number of breaks and that the breaks are common to the parameters and group memberships of all units. Our method is different as it allows for cluster switches of individual units at any point in time.

The remainder of this paper is set up as follows. In Section 2 we introduce the methodology. Section 3 considers a simplified setting where we study misclassification probabilities and optimal penalty parameters analytically. Section 4 studies the new approach in a controlled environment and shows reductions in overall misclassification rates in line with our analytical results. Section 5 discusses the empirical application, while Section 6 concludes. An appendix provides proofs and further technical and empirical results.

2 Methodology

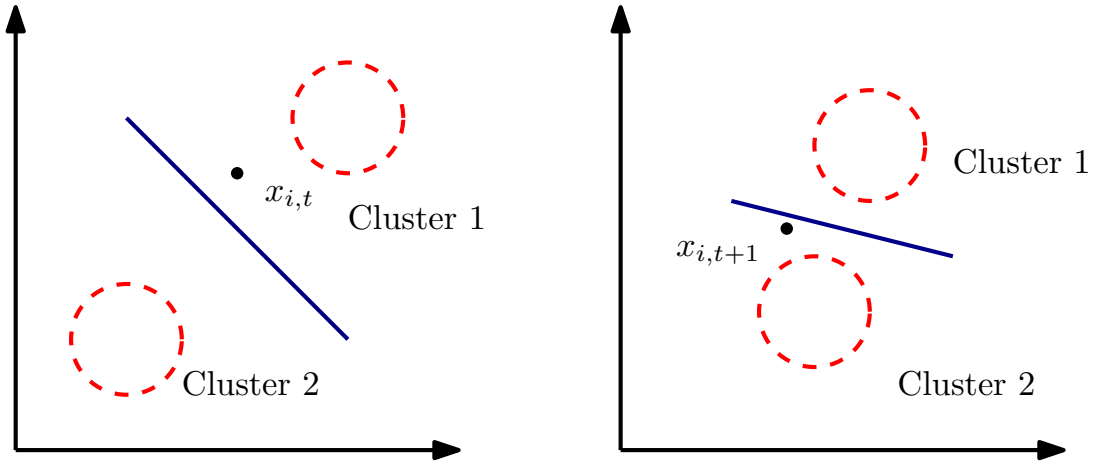
In this section, we first introduce our robust clustering methodology. Next, we explain how we link cluster identities over time, which is a crucial step in our method. Finally, we provide data-driven ways to select the shrinkage penalty parameter in our approach.

2.1 Penalized cross-sectional clustering

Consider a panel of multivariate financial data, with $x_{i,t} \in \mathbb{R}^{D\times 1}$ denoting a vector of observed characteristics for unit $i = 1, \ldots, N$ at time $t = 1, \ldots, T$. Our goal is to assign each unit i to a peer group of similar units at each point in time t. An example of such a situation is the monitoring of business models in the financial industry by a prudential supervisor as in Custodio João et al. (2022). In the realistic setting of changing market conditions, technological advances, and shifts in regulatory requirements, we expect that some firms may move to a different group or business model at some point in time. However, switching from group A to group B at one point in time, only to switch back from B to A in the following period, is unrealistic in many situations that involve long-term strategies. A suitable clustering method should therefore mitigate excessive cluster switches.

To illustrate this, consider an example with $D = 2$ features in Figure 1. Assume we cluster each cross-section t separately into two clusters by, for instance, a k-means approach. Units are then assigned to the cluster with the closest cluster center. This divides the space in two regions. If an observation $x_{i,t}$ at time t is close to the border that separates the clusters, as in the left-hand panel, even a small disturbance to its position might shift it to the other cluster. A second switch might then occur if it is subject to another small and roughly opposite disturbance in the next period, and so on. We would observe short-lived cluster switches, or "flickering", caused by little actual movement. Such flickering might not be economically meaningful, and therefore undesirable.

The approach presented in this paper takes the cross-section at time t and combines it with the $t-1$ cluster assignments to produce sticky assignments over consecutive cross-



(a) time t (b) time $t+1$

Figure 1: k-means clustering at two consecutive times. The red circles represent the location of the cluster centers. The blue line separates the clusters and is halfway between both cluster centers.

sections. For instance, if a unit is assigned to cluster A at time $t-1$, we first shrink that unit's observation at time t towards the mean of cluster A at time t before re-classifying it. To solve the arbitrary labeling of clusters over different cross-sections, we propose a mapping procedure based on the maximum overlap between cluster membership: for instance, if 90% of the units in a particular cluster at time t have the same identity as what was called 'cluster A' at time $t-1$, then we label that cluster also 'cluster A' at time t. The precise mapping procedure is explained in detail in Section 2.2.

We introduce some notation and present the formal algorithm, which is summarized in Algorithm 1. Let $h_{i,t}$ denote the cluster assignment of unit i at time t, such that $h_t = (h_{1,t}, \ldots, h_{N,t})'$ denotes the $N \times 1$ vector of all cluster assignments for cross-section t. We now start at time $t = 1$ with a standard cross-sectional clustering algorithm and cluster selection criterion to obtain the number of clusters K_t and the cluster identities h_t at $t = 1$. Next we move to $t = 2$ and run a clustering algorithm to obtain a *candidate* set of cluster assignments $\tilde{h}_t$. Using the mapping methodology M of Section 2.2, we relabel the cluster identities in $\tilde{h}_t$ to $\tilde{h}'_t = M(h_{t-1}, \tilde{h}_t)$, such that the identities in h_{t-1} and $\tilde{h}'_t$ are comparable. Based on $\tilde{h}'_t$, we compute the *current* (candidate) location of each of the *previous* clusters in h_{t-1}, except for the clusters that were discontinued. For example, if we estimate the cluster location by its mean, the current (candidate) location

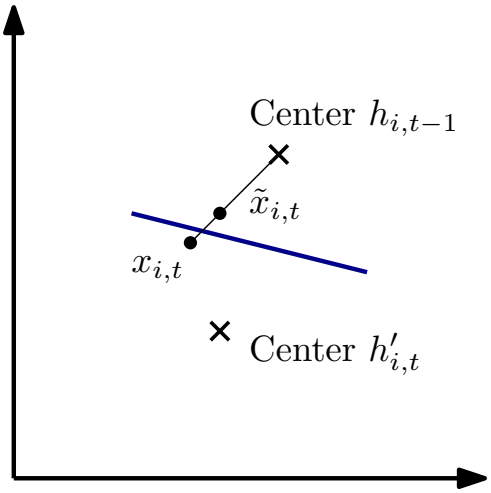


Figure 2: We can interpret $\tilde{x}_{i,t}$ as an artificial position lying an ε fraction of the way from the current position $x_{i,t}$ and the center of its last cluster $h_{i,t-1}$

of unit i's previous cluster can be estimated by $c(h_{i,t-1}, \tilde{h}'_t) = (\#P_i)^{-1} \sum_{j \in P_i} x_{j,t}$, where $P_i = \{j \mid \tilde{h}'_{j,t} = h_{i,t-1}\}$ is the set of current (candidate) units that are in the same cluster now that unit i was in a period ago, and $\#P_i$ denotes the number of elements in P_i. If the number of elements in P_i is positive, we then shrink $x_{i,t}$ towards the current location of its previous cluster. We do so by defining

$$\tilde{x}_{i,t} = (1 - \varepsilon) \cdot x_{i,t} + \varepsilon \cdot c(h_{i,t-1}, \tilde{h}'_t), \tag{1}$$

where ε is a fixed penalty parameter in the unit interval. The effect can be seen in Figure 2.

Using the shrunk observations $\tilde{x}_{i,t}$, we run a second pass of the cluster assignments as

$$h_{i,t} = \mathbb{1}_{i,t} \cdot \tilde{h}'_{i,t} + (1 - \mathbb{1}_{i,t}) \cdot h_{i,t-1}, \tag{2}$$

$$\mathbb{1}_{i,t} = \begin{cases} 1 & \text{if} \quad \#P_i = 0 \quad \text{or } d\big(\tilde{x}_{i,t}, c(h'_{i,t}, \tilde{h}'_t)\big) < d\big(\tilde{x}_{i,t}, c(h_{i,t-1}, \tilde{h}'_t)\big), \\ 0 & \text{else,} \end{cases} \tag{3}$$

where d is a distance measure. In words: if the shrunk observation $\tilde{x}_{i,t}$ is closer to the new candidate cluster, or if the old cluster is discontinued, the unit switches to the new

Algorithm 1: Dynamic clustering with shrinkage

input : The data, the maximum number of clusters K^{max} per cross-section, a shrinkage parameter ε.

output: T vectors of assignments h_t.

for $t \in [T]$ **:**

  for $\widetilde{K} \in \{2, 3, \ldots, K^{max}\}$ **:**

    Run clustering algorithm; obtain candidate cluster assignments $\tilde{h}_t$

    if $t > 1$ **then**

      $\tilde{h}'_t \leftarrow M(h_{t-1}, \tilde{h}_t)$

      compute new locations of previous clusters: $c(h_{i,t-1}, \tilde{h}'_t)$

      shrink observations to current means of previous clusters: $\tilde{x}_{i,t}$

      $h^{(\widetilde{K})} \leftarrow$ re-assign shrunk observations to clusters

    else

      $h^{(\widetilde{K})} = \tilde{h}_1$

    $s_t^{\widetilde{K}} \leftarrow$ compute silhouette index for this cross-section based on current cluster assignments $h^{(\widetilde{K})}$

  $K_t \leftarrow \arg\max_{\widetilde{K}} s_t^{\widetilde{K}}$; select number of clusters in cross-section t

  $h_t \leftarrow h^{(K_t)}$; store final assignments for cross-section t

cluster. Otherwise, the unit remains in the old cluster.[1] The shrinkage of the observation towards the current location of the previous cluster ensures that cluster switches become less likely. If ε equals zero, there is no shrinkage and units can switch cluster identity freely from one cross-section to the next. The steps are repeated for all cross-sections $1, \ldots, T$, including a step to determine the number of clusters in each cross-section. The complete algorithm is summarized in Algorithm 1.

It is important to note here that we have been silent thus far about which clustering algorithm is used, which distance measure d, and which measure of cluster centroid c. This means that the current shrinkage technique can be applied in a wide variety of settings. Any cross-sectional clustering algorithm that produces a distance measure can be adapted in the above way to feature stickiness. For example, in graph-based algorithms such as in Zahn (1971) or Grundmann et al. (2010), we can shrink the weight of edges connecting points that belonged to the same cluster at $t-1$. For simplicity, we use k-means in our simulations and empirical application, but other methods are possible if the data calls

[1] The procedure of candidate clustering, mapping, shrinking, and reassignment, could be iterated if desired. Also note that the approach could, in principle, be extended from the current hard clustering assignment procedure to a soft clustering assignment.

for more exotic cluster shapes. We can be similarly flexible with regard to the choices of distance measures d and centroids c. For instance, if distances are Mahalanobis-based, we can choose to pool across all cross-sections to compute (cluster) covariance matrices, or alternatively compute such matrices per cross-section, thus allowing for heteroskedasticity.

To select the number of clusters in each cross-section in Algorithm 1, we use the silhouette index of Rousseeuw (1987). Like other cluster selection criteria, it favors homogeneity of units within each cluster as well as heterogeneity between clusters; better scores on either dimension result in higher values of the index. We pick the number of clusters K_t that maximizes the average silhouette index. The silhouette of point i at time t for a given K_t is

$$s_{it} = \frac{b(x_{i,t}) - a(x_{i,t})}{\max\{a(x_{i,t}), b(x_{i,t})\}}, \qquad a(x_{i,t}) = d(x_{i,t}, C_{h_{i,t},t}), \qquad b(x_{i,t}) = \min_{k \neq h_{i,t}} d(x_{i,t}, C_{k,t}), \quad (4)$$

where $a(x_{i,t})$ is the average distance from point i to other points in its own cluster $C_{h_{i,t},t}$, and $b(x_{i,t})$ is the average distance from point i to the points in the nearest other cluster. Following Rousseeuw (1987), we set $s(x_{i,t}) = 0$ if $C_{h_{i,t},t}$ only contains unit i. Intuitively, the average silhouette index

$$s_t = \frac{1}{N} \sum_{i=1}^{N} s_{it} \quad (5)$$

measures how tightly the observations are clustered around the cluster mean (when $a(x_{i,t})$ is low on average) and how separate the clusters are from each other (when $b(x_{i,t})$ is high on average). This makes it a useful measure of fit, which we adapt in Section 2.3 to obtain a data-driven way to select the shrinkage parameter ε.

2.2 Mapping

Standard cross-sectional clustering algorithms produce arbitrary cluster labels that have no relation to the labels assigned in previous cross-sections. This complicates the identification of the current location of an observation's previous cluster. To remedy this, we need to find a correspondence between the labels at $t-1$ and new candidate labels at

time t. We do so by looking at the overlap of every two clusters at consecutive times, as in Kalnis et al. (2005). To illustrate, consider a setting where at time t the cross-sectional clustering algorithm produces clusters A and B, while at time $t+1$ it produces clusters labeled C and D. If all units that belong to cluster A at time t also belong to cluster D at time $t+1$, and all units that belong to cluster B at time t belong to cluster C at $t+1$, then the most natural correspondence is to assign the same label to A and D, and similarly to B and C. Following Oliveira and Gama (2010), we refer to this procedure as *mapping*.

To generalize this idea to the less obvious case where there are switches, we form a contingency matrix where the elements in row i and column j represents how many points were assigned to cluster i at time t and to cluster j at time $t+1$. We can then formalize the idea of maximizing overlap between clusters at different times as maximizing the trace of this matrix with respect to the ordering of the columns. For example, if both periods have two clusters and the maximum is attained when the contingency matrix is formed with the column corresponding to cluster D on the left and the one for cluster C on the right, then cluster D at $t+1$ maps to cluster A (row 1) at t, and so on, as is the case in the example below:

$$\begin{array}{c} \\ A: \\ B: \end{array} \begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \mathrm{tr}\begin{pmatrix} 3 & 2 \\ 5 & 1 \end{pmatrix} = 4 \end{array} \quad \rightarrow \quad \begin{array}{c} \\ A: \\ B: \end{array} \begin{array}{c} \begin{array}{cc} D & C \end{array} \\ \mathrm{tr}\begin{pmatrix} 2 & 3 \\ 1 & 5 \end{pmatrix} = 7. \end{array}$$

This formal problem can be written as

$$\max_{P} \mathrm{tr}(C_t^* P), \tag{6}$$

where C_t^* is the contingency matrix from time t to $t+1$, and P is a permutation matrix. The optimal P can easily be interpreted: cluster i at $t+1$ maps to cluster j at t if $P_{i,j} = 1$. For a small number of clusters at $t+1$ (7 or 8, say), (6) can be easily solved by exhaustive search. For larger numbers of clusters, an efficient algorithm has been developed, known

as the Hungarian algorithm (Kuhn, 1955).

An extension of the above method to the situation where the number of clusters increases or decreases over time can be defined as follows. The contingency matrix then becomes rectangular, so it is no longer possible to compute its trace. This is solved by maximizing the trace of the largest square matrix inside it by switching the columns of the rectangular matrix. That is, the extra clusters' overlap will not go into the objective function. We can still formulate the problem as in equation (6) if we augment C_t^* with a matrix of zeroes such that the resulting matrix is square, i.e.,

$$C_t^{**} = \begin{pmatrix} C_t^* \\ \mathbf{O}_{m-n\times m} \end{pmatrix}, \qquad C_t^{**} = \begin{pmatrix} C_t^* & \mathbf{O}_{n\times n-m} \end{pmatrix}, \tag{7}$$

for the case $n < m$ and $n > m$, respectively, where $\mathbf{O}_{a\times b}$ is the matrix of zeroes of dimension $a \times b$. The problem can then again be written as (6), with C_t^{**} taking the place of C_t^*. This formulation is equivalent to stating that the extra clusters exist in both time steps but have no members in one of them (as Frühwirth-Schnatter and Malsiner-Walli (2019) do in their paper). Therefore, this extended problem can still be solved by the Hungarian algorithm. It does not, however, provide a solution for ties, i.e. when two relabellings of the clusters at $t+1$ provide the same trace $\text{tr}(C_t^* P)$. Luckily, such situations are empirically exceedingly rare. Still, in such rare cases ties can be broken in a variety of ways, for instance by considering the overlap with the cross-section at $t-2$, or by using the correspondence with the closest cluster means.

2.3 Selection of the shrinkage parameter

In this section, we propose a modification of the silhouette index to set the shrinkage parameter ε in (1). In Section 4 we benchmark this statistic against the cross-validation approach of Fu and Perry (2020), which we also briefly introduce here. The latter, however, turns out to work less well.

Our aim is to reduce misclassification rates of observations to clusters. As clustering

is an unsupervised learning technique, such misclassification can only be studied in a controlled setting, as we do in Sections 3 and 4. For real data, misclassification cannot be measured and we look instead at measures of cluster fit based on the silhouette index. Here a trade-off has to be made between the best fit on the one hand, and stability of cluster assignments (no undue flickering) on the other. As flickering is a highly transitory phenomenon, we take advantage of this fact to inform our choice of ε. Specifically, we look for values of ε that have a large effect on bringing down the number of switches, but only a modest effect on the overall clustering fit as measured by the silhouette index.

To aggregate the silhouette index across cross-sections, we use the Gini-weighted average version as in David (1968), re-scaled by N:

$$G_t = \frac{\sum_{k=1}^{K_t}(2k - K_t - 1) \cdot \#P_{k:K_t}}{K_t \cdot N} = \frac{\sum_{i=1}^{K_t}\sum_{j=1}^{K_t}\left|\#P_i - \#P_j\right|}{2K_t \cdot N},$$
$$GWS = \sum_{t=1}^{T}(1 - G_t) \cdot s_t,$$

where $\#P_{k:K_t}$ is the number of units in the k-th smallest cluster, and s_t denotes the silhouette index of cross-section t. The second expression for G_t clearly shows that G_t equals zero when the clusters have homogeneous sizes, and increases as inequality in cluster size increases.

An important reason for choosing the GWS over other measures is that it penalizes clusters with a single outlying observation. Rousseeuw (1987) already notes that the simple average silhouette could be vulnerable to outliers: "*a situation where the data set contains one far outlier is also an example of a strong clustering structure. Indeed, when the outlier is far enough, the other data look like a tight cluster by comparison.*" By applying the Gini weights rather than computing a simple average, we avoid picking cluster numbers that result in single (outlying) observation clusters. This is also in line with our objective to reduce flickering: we want to discourage the short-lived birth and death of small, isolated clusters from one cross-section to the next. The GWS statistic is easy to compute and a direct by-product of Algorithm 1.[2]

[2]The silhouette index is available at the level of each unit i, (see equation (4)), on average across i

To benchmark the Gini weighted silhouette (GWS) index, we also compute the cross-validation statistic for clustering proposed by Fu and Perry (2020). In their paper, Fu and Perry use cross-validation to determine the optimal number of clusters in a cross-sectional clustering problem. We, instead, use their approach to set the shrinkage parameter ε. Following Fu and Perry (2020), we first randomly split the units of our dataset in (three) equal groups, as well as the variables (in two groups). Next we build six folds out of these groups in the following way. One of the three groups of units is assigned to be the training set, while the other two are the test set. Also, one group of variables is taken as predictor variables (X_t^{tr} and X_t^{te} for the train and test sample, respectively). The other variables are called response variables (Y_t^{tr} and Y_t^{te}). In each data fold, we apply four steps to reach a measure of cross-validation error. First, we *cluster* Y_t^{tr} using our shrinkage methodology to obtain labels c_t^{tr} and corresponding cluster means $\mu_{t,k}^{tr,Y}$ for the training response variables and the $k = 1, \ldots, K_t$ clusters. We also cluster the observations in Y_t^{te} using the same shrinkage approach, but based on the already estimated cluster means $\mu_{t,k}^{tr,Y}$ and the same number of clusters K_t. This gives us cluster labels c_t^{te}. We treat the labels c_t^{tr} and c_t^{te} as observed 'pseudo-labels' in the next step. Next, we perform a *classification* step on X_t^{te}. Though in principle any classification model could be used, we follow Fu and Perry (2020) and use a simple classifier that estimates cluster means $\mu_{t,k}^{tr,X}$ of X_t^{tr} based on the assignments c_t^{tr}. We then *predict* cluster assignments $\hat{c}_t^{te}$ for X_t^{te} by assigning each observation in X_t^{te} to the cluster with the closest mean $\mu_{t,k}^{tr,X}$. Our cross-validation error is then $||Y_t^{te} - \mu_{t,\hat{c}_t^{te}}^{te,Y}||^2$, i.e., the prediction error for the response variable in the testing sample based on the predictor variables' classification model. The squared cross-validation errors are averaged over all observations and all folds, and subsequently minimized to compute the optimal shrinkage parameter ε.

at any point in time t, and for the entire data. At the unit level, it compares the closest fit of unit i to its second-best cluster alternative at time t, taking into account all other possible cluster allocations. As a result, s_{it} can play a role similar to the role that the cluster probabilities $\tau_{ij,1:T}$ play in Lucas et al. (2019), and that the filtered cluster probabilities $\tau_{ij,t|t}$ play in Custodio João et al. (2022).

3 Analytical results for a simplified model

This section presents a simplified clustering model that allows us to investigate analytically under which conditions a higher shrinkage parameter ε improves correct classification rates. We consider a univariate data generating process, where the observation x_t depends on its cluster center c_t, according to

$$x_t = c_t + \eta_t,$$

where η_t has cdf $F(\eta_t)$ with zero mean and variance σ^2. The clusters are labeled based on their cluster center c_t, which we normalize to 0 and 1, i.e., $c_t \in \{0, 1\}$. The similarity of the clusters is then fully determined by the cdf $F(\cdot)$. If the support of $F(\eta_t)$ is highly concentrated around $\eta_t = 0$, then the clusters are well-separated. If, by contrast, the support of $F(\eta_t)$ is widespread around zero, then the clusters are very similar and difficult to distinguish.

We allow for potential switching of cluster membership by assuming that c_t follows a Markov chain with transition probability p. The repeated k-means clustering procedure in this setting then corresponds to classifying x_t to either the cluster with center 0 ($\hat{c}_t = 0$) or 1 ($\hat{c}_t = 1$), regardless of the previous cluster assignment $\hat{c}_{t-1}$, i.e., using $\varepsilon = 0$ and basing the assignment on $\hat{c}_t = \arg\min_{c \in \{0,1\}} |x_t - c|$.

The penalized clustering methodology introduced in Section 2 relies on the previous assignment $\hat{c}_{t-1}$. It can be written as

$$\hat{c}_t^{\varepsilon} = \hat{c}^{\varepsilon}(x_t | \hat{c}_{t-1}) = \begin{cases} 1 & \text{if} \quad x_t(1-\varepsilon) + \varepsilon \hat{c}_{t-1} > 1/2, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

We call this the ε-classifier. The repeated k-means procedure is a special case of this approach for $\varepsilon = 0$.

Using the above set-up, we can analytically derive the probability of misclassification $\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_t)$. In our first result, we define the one-step-ahead misclassification rate as

the misclassification probability at t given perfect information about the true cluster membership at $t-1$.

Proposition 1. *Given c_{t-1}, the one-step-ahead misclassification probability of the ε-classifier is*

$$\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_t | c_{t-1}) = F\left(\frac{\varepsilon - 1/2}{1-\varepsilon}\right) p + F\left(\frac{-1/2}{1-\varepsilon}\right)(1-p), \tag{9}$$

where F denotes the cdf of η_t.

All proofs can be found in Appendix A. A few features of (9) are worth noting. The k-means error ($\varepsilon = 0$) simplifies to $\mathbb{P}(\hat{c}_t \neq c_t | c_{t-1}) = F(-1/2)$, which is insensitive to p. On the other extreme, as $\varepsilon \to 1$, the error approaches p. Figure 3 plots the misclassification probability for the more interesting intermediate values of ε using a normal distribution $\mathrm{N}(0, \sigma^2)$ for $F(\eta_t)$. The minimum of each curve is marked by a dot. In most cases the minimum classification error is obtained at some intermediate value of ε. The reduction in classification errors is larger for smaller values of p, i.e., situations where there is infrequent switching and where time $t-1$ information is most informative about the cluster identity at time t. Improvements are also larger for more cluster similarity (high σ^2). Only for $p = 0.5$ the position at $t-1$ does not bear any information for the cluster assignment at t, and there is no benefit in introducing membership persistence via $\varepsilon > 0$. Later in Section 4, we will see that the current analytical results bear close resemblance to the simulation results.

Using Proposition 1, we can analytically characterize the optimal value of ε that minimizes the misclassification rate.

Proposition 2. *If $\eta_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, the value ε^* which minimizes the misclassification rate (9) for $0 < p < \frac{1}{2}$ is*

$$\varepsilon^* = \frac{2\sigma^2 \log\left(\frac{p}{1-p}\right)}{2\sigma^2 \log\left(\frac{p}{1-p}\right) - 1}. \tag{10}$$

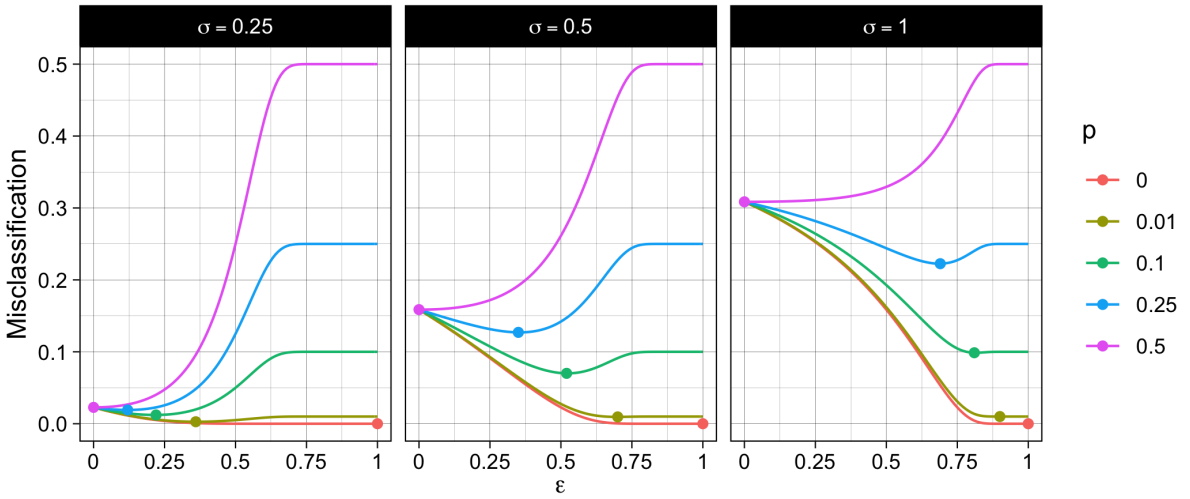


Figure 3: Plot of the one-step-ahead misclassification probability (9)
The dot on each curve indicates the lowest misclassification rate. The figure is based on $\eta_t \sim \mathrm{N}(0, \sigma^2)$, $x_t = c_t + \eta_t$, and $c_t \in \{0, 1\}$ a Markov chain with switching probability p. For $\sigma = 0.25$ the clusters are well-separated, while for $\sigma = 1$ the clusters largely overlap.

Proposition 2 confirms what can be seen from Figure 3. Higher levels of noise and lower switching probabilities p push ε^* upwards, implying that higher shrinkage and thus more persistence in the classifier is optimal in such cases. In most cases of empirical interest, switching is present but infrequent ($0 < p < 0.5$), leading to strictly positive values of ε^*.

We can extend Proposition 1 to the case where x_t exhibits mean-reverting dynamics in addition to a Markov switching center.

Corollary 1. *If x_t follows the dynamics $x_t = c_t + \beta(x_{t-1} - c_t) + \eta_t$, then the one-step-ahead probability of error of the ε-classifier is*

$$\begin{aligned}\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_t | c_{t-1}) =& F\left(\frac{\varepsilon - 1/2}{1-\varepsilon} + \beta(x_{t-1}(2c_{t-1} - 1) + 1 - c_{t-1})\right) p \\ &+ F\left(\frac{-1/2}{1-\varepsilon} + \beta(x_{t-1}(1 - 2c_{t-1}) + c_{t-1})\right)(1-p),\end{aligned} \tag{11}$$

where F denotes the cdf of η_t.

We note that the introduction of the term $\beta(x_{t-1} - c_t)$ does not change the main features of the misclassification probability (11) when compared to (9). In particular, the concavity of the misclassification rate is still present, as is the strictly positive optimal

value of ε; see Figure B.1 in the Appendix B.

Proposition 1 assumes that the true past cluster mean c_{t-1} is known. This is admittedly unrealistic. To arrive at an unconditional misclassification rate, and a corresponding optimal shrinkage parameter ε^*, we propagate the classification process n steps ahead to derive the misclassification rate $\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_t | c_{t-n})$. Let the conditional correct classification probabilities at time t be $q_{i,t} = \mathbb{P}(\hat{c}_t^{\varepsilon} = i | c_t = i)$ for $i = 0, 1$, and let $q_t = (q_{0,t}, q_{1,t})'$. Also define the marginal probabilities of the true states $\pi_{i,t} = \mathbb{P}(c_t = i)$ with $\pi_t = (\pi_{0,t}, \pi_{1,t})'$, such that the probability of correct classification can be written as $\pi_t' q_t = q_{0,t}\pi_{0,t} + q_{1,t}\pi_{1,t}$. Then the following proposition gives the recursion for q_{t+1}.

Proposition 3. *The conditional correct classification probabilities q_t follow the recursion*

$$q_{t+1} = \begin{pmatrix} \frac{z_{00}\cdot(1-p)\cdot\pi_{0,t}}{\pi_{0,t+1}} - \frac{z_{10}\cdot(1-p)\cdot\pi_{0,t}}{\pi_{0,t+1}} & \frac{z_{10}\cdot p\cdot\pi_{1,t}}{\pi_{0,t+1}} - \frac{z_{00}\cdot p\cdot\pi_{1,t}}{\pi_{0,t+1}} \\ \frac{z_{01}\cdot p\cdot\pi_{0,t}}{\pi_{1,t+1}} - \frac{z_{11}\cdot p\cdot\pi_{0,t}}{\pi_{1,t+1}} & \frac{z_{11}\cdot(1-p)\cdot\pi_{1,t}}{\pi_{1,t+1}} - \frac{z_{01}\cdot(1-p)\cdot\pi_{1,t}}{\pi_{1,t+1}} \end{pmatrix} \cdot q_t \tag{12}$$

$$+ \begin{pmatrix} \frac{z_{00}\cdot p\cdot\pi_{1,t}}{\pi_{0,t+1}} + \frac{z_{10}\cdot(1-p)\cdot\pi_{0,t}}{\pi_{0,t+1}} \\ \frac{z_{01}\cdot(1-p)\cdot\pi_{1,t}}{\pi_{1,t+1}} + \frac{z_{11}\cdot p\cdot\pi_{0,t}}{\pi_{1,t+1}}, \end{pmatrix}$$

where

$$z_{i0} = F\left(\frac{1/2 - i\cdot\varepsilon}{1-\varepsilon}\right), \quad z_{i1} = 1 - F\left(\frac{1/2 - i\cdot\varepsilon}{1-\varepsilon} - 1\right).$$

We note that Proposition 1 is a special case of (12) by taking $q_t = (1, 1)'$. Other values can be chosen to reflect the uncertainty in the first step. In particular, we can use the output of the first step as input for a second step to obtain the 2-step-head error rate. This process can be repeated n steps. By iterating further and further, we can study whether introducing persistence in clustering ($\varepsilon > 0$) has a lasting benefit, whatever the initialization used. The following two corollaries present the results for $n \to \infty$, establishing that some strictly positive shrinkage parameter is generally optimal even if no information is available about the previous cluster label c_{t-1}. The result is established for our case of a symmetric Markov chain for c_t, where $\lim_{t\to\infty} \pi_{i,t} = 0.5$ for

$p > 0$. Derivations for an asymmetric Markov chain are very similar.

Corollary 2. *The limiting probabilities of correct classification q for a symmetric Markov chain $\mathbb{P}(c_t = 1|c_{t-1} = 0) = \mathbb{P}(c_t = 0|c_{t-1} = 1) = p$ are*

$$q = \begin{pmatrix} 1-(1-p)(z_{00}-z_{10}) & p\cdot(z_{00}-z_{10}) \\ p\cdot(z_{11}-z_{01}) & 1-(1-p)(z_{11}-z_{01}) \end{pmatrix}^{-1} \times \begin{pmatrix} z_{10}+p\cdot(z_{00}-z_{10}) \\ z_{01}+p\cdot(z_{11}-z_{01}) \end{pmatrix}.$$

The corresponding limiting misclassification probability is

$$\lim_{t\to\infty} \mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_t) = 1 - \frac{1}{2}\frac{z_{01}(1-\tilde{z}_{00}) + z_{10}(1-\tilde{z}_{11}) + p(\tilde{z}_{00}+\tilde{z}_{11}-2\tilde{z}_{11}\tilde{z}_{00})}{1-(1-p)(\tilde{z}_{00}+\tilde{z}_{11}) + (1-2p)\tilde{z}_{00}\tilde{z}_{11}}, \tag{13}$$

where $\tilde{z}_{00} = z_{00} - z_{10}$ and $\tilde{z}_{11} = z_{11} - z_{01}$.

Corollary 3. *Let $f(\eta_t)$ be the pdf of η_t, corresponding to the cdf $F(\eta_t)$. Then under the same conditions as Corollary 2, the derivative of the limiting misclassification probability at $\varepsilon = 0$ is given by*

$$\left.\frac{\partial\, \frac{1}{2}(1-(1\ ,\ 1)\ q)}{\partial\varepsilon}\right|_{\varepsilon=0} = \tfrac{1}{4}\left(f(\tfrac{1}{2}) - f(-\tfrac{1}{2})\right) + \tfrac{1}{2}p\left(f(-\tfrac{1}{2})F(\tfrac{1}{2}) - f(\tfrac{1}{2})F(-\tfrac{1}{2})\right) + \tfrac{1}{2}(p-1)\left(f(\tfrac{1}{2})F(\tfrac{1}{2}) - f(-\tfrac{1}{2})F(-\tfrac{1}{2})\right).$$

If the pdf f is symmetric around zero, this expression simplifies to

$$\left.\frac{\partial\, \frac{1}{2}(1-(1\ ,\ 1)\ q)}{\partial\varepsilon}\right|_{\varepsilon=0} = -\tfrac{1}{2}(1-2p)\ f(\tfrac{1}{2})\left(2F(\tfrac{1}{2}) - 1\right),$$

which is negative for $p < 0.5$.

Figure 4 plots the n-step-ahead misclassification rate from Proposition 3 and its limit from Corollary 2 for $n \in \{1, 5, \infty\}$, and $\eta_t \sim \mathrm{N}(0, 0.5^2)$. Introducing clustering persistence clearly pays off both in the short and the long term. The minimum misclassification rate is reached at some $\varepsilon > 0$ as long as $p < 0.5$. This follows from the derivative of the misclassification rate at the origin $\varepsilon = 0$ in Corollary 3, which is negative for any $p < 0.5$

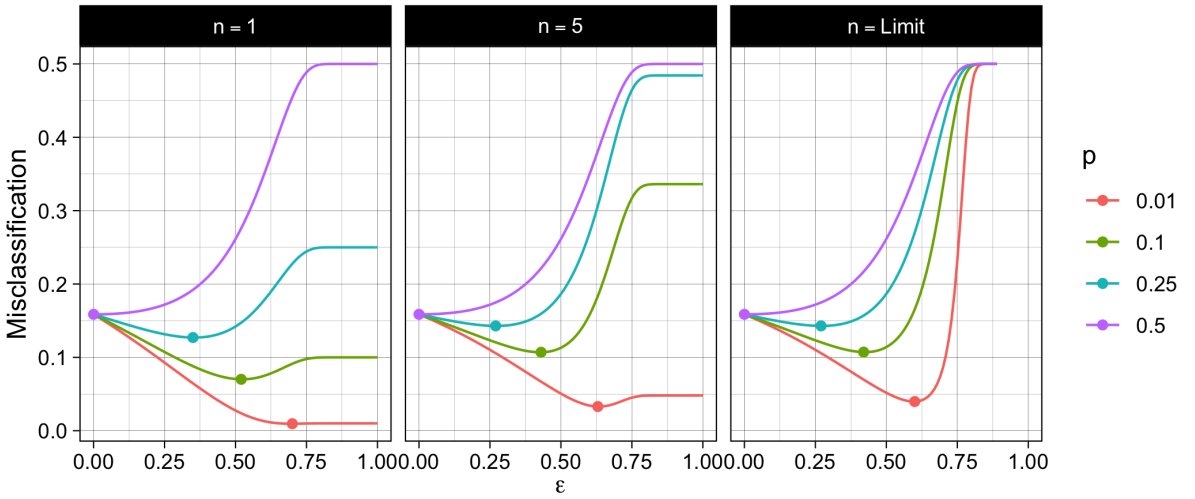


Figure 4: The n-step-ahead misclassification rate $\pi_t' q_t$ for $n = 1, 5$ and $n \to \infty$ using (12) and $\sigma = 0.5$.

and any distribution F of η_t that is symmetric around zero. We also note that the drop in the misclassification rate remains substantial for the limiting case $n \to \infty$ for $p \leq 0.1$. Finally, all these results also align with our simulations results in Section 4: under moderate switching, the error rate is concave in ε, and more so if clusters are less well separated (i.e., higher σ). This helps to recognize situations for which it is advisable to allow for cluster switching over time, balanced with the shrinkage approach proposed in this paper.

4 Simulation study

In this section, we investigate the ability of our method to assign units to their respective clusters at each point in time. All simulations are done using a six-dimensional Gaussian distribution ($D = 6$) to allow for at least three variables in each fold of the cross-validation approach of Fu and Perry (2020). The different cluster centers are drawn randomly from the vertices of a six-dimensional unit hypercube. In the baseline simulation setting, the cluster covariance matrices are set equal to the identity matrix. Unit variances for cluster centers on the vertices of a unit cube imply that there is substantial cluster overlap and thus substantial misclassification risk. We therefore also consider a second setting with variances equal to 0.5 for each component. Throughout the simulations, the true number

of clusters is fixed at two ($K = 2$).

At each time, observations are drawn from their current cluster distribution. Units switch clusters from time t to $t + 1$ with probability p, where we vary p from 0 to 0.25 across different designs. In all settings we use $T = 20$ time points, $N = 120$ units, and 100 simulations runs. As our first-pass cross-sectional clustering algorithm we choose a simple k-means approach, although, as stated before, our approach can also accommodate other cross-sectional clustering methods, distance definitions, and centroid measures.

The baseline simulation results are shown in Figure 5. Overall, the shape of the misclassification curve closely aligns with our analytical results in Figures 9 and 12 in Section 3. At low levels of switching in the DGP, i.e. $p \in \{0, 0.01, 0.1\}$, our method with positive ε improves on the repeated k-means case ($\varepsilon = 0$). Moreover, there are clearly optimal values for ε in the misclassification plot (upper left panel). Setting ε to these optimal values leads to reductions of misclassification errors from 16% down to 9% for both $p = 0$ and $p = 0.01$. The shrinkage approach performs worse than repeated cross-sectional clustering only for highly-frequent actual switching ($p = 0.25$). We do not expect this to be a major problem in practice, as our model is primarily intended for dynamic settings with only occasional switches and substantial persistence in cluster membership.

Without knowing the true classifications, it is still striking that the Gini-weighted Silhouette index peaks at about the optimal ε (lower-right panel), while the cross-validation (CV) error flattens out around the same point (upper-right panel). The switching rate (lower-left panel) combined with the Gini-weighted silhouette index shows exactly what the approach seeks to achieve – a drastic reduction in the number of cluster switches (lower-left), without sacrificing the fit in terms of the silhouette index (lower-right). Increasing ε avoids frequent reclassification of observations on the borderline between clusters. Such observations only marginally affect the silhouette index as it takes the distances of the observations to the nearest clusters into account. It does, however, bring down the switching rate considerably. This may help setting the value of ε in empirical applications: we are looking for values of ε that reduce the switching rate, without considerable decreases in the silhouette index.

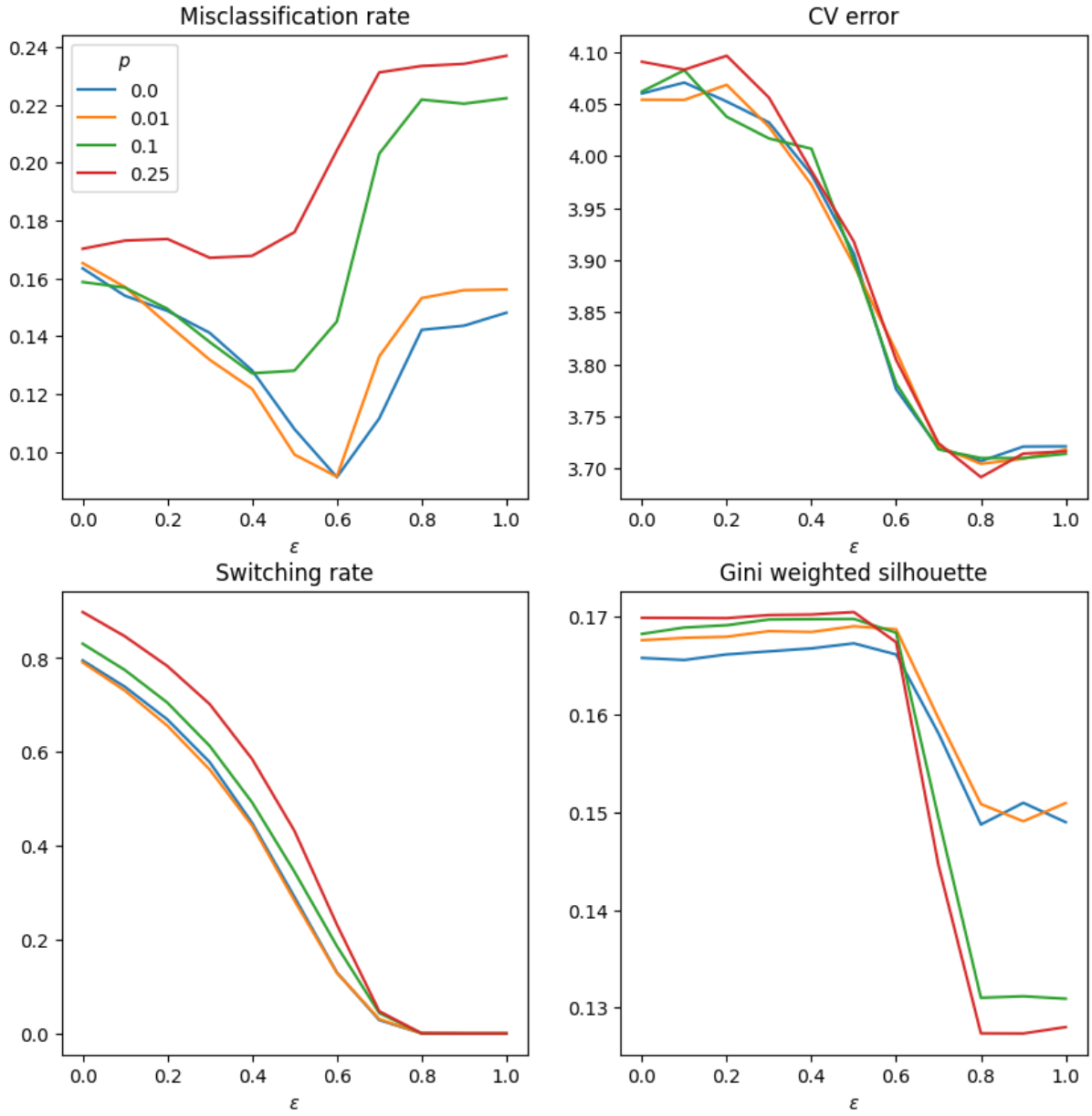


Figure 5: Simulation results for four values of p. Baseline setting.

We emphasize that our baseline setting implies a major challenge to any clustering algorithm, owing to the substantial cluster overlap. In a setting with lower variances, such as Figure B.2 in Appendix B, we find much smaller misclassification rates, while still achieving reductions in misclassification rates of about 67% (from around 7.5% to around 2.5%) for $p = 0.00$ and 0.01. Again, the pronounced concavity and minimum of the misclassification reflect the theoretical results in Section 3.

Figure 5 also suggests that the cross-validation error approach to select ε works less well. Cross-validation errors appear lowest for high values of ε that yield too much

persistence in cluster membership. If ε were set based on this criterion, misclassification rates would be higher than those associated to the Gini-weighted silhouette approach. We therefore prefer the latter over the former in our empirical work in Section 5.

To see the effect of choosing the number of clusters, we extend the previous simulation setup by also letting the algorithm choose the number of clusters K_t in each cross-section. We vary the number of clusters in the model from 2 to 4, whereas the true number of clusters is always 2. The results are presented in Figures B.3 and B.4 in Appendix B. The case of an unknown number of clusters, combined with a large cluster overlap, poses a substantial challenge for any clustering method. Misclassification rates are high throughout, and only for $p = 0.00, 0.01$ we observe a clear dependence on the shrinkage parameter ε. For those two cases, the reduction in misclassification is substantial, at more than 15 percentage points when the optimal penalty parameter is chosen. The Gini-weighted silhouette index points to values of ε between 0.3 and 0.6, where the sharpest declines in the Gini-weighted silhouette index occurs. These values appear slightly below the optimal values for misclassification at around $\varepsilon = 0.6$. The CV error, by contrast, seems to flatten out around too high a value of around $\varepsilon = 0.8$, and thus again exhibits a worse behavior. The picture is even clearer if we bring down the error variances to 0.5, reducing cluster overlap. For low values of p, the Gini-weighted silhouette index now decreases sharply after the optimal (from a misclassification perspective) value of ε has been reached. This allows us to cut misclassification by close to 50% in a data-driven way without sacrificing much of the fit in terms of the Gini-weighted silhouette index. By contrast, applying the cross-validation-based approach again results in too high values of ε and may therefore miss important aspects in the dynamics of the data.

Finally, to benchmark our new clustering approach, we compare it to three versions of Ward's hierarchical clustering. The first approach (Ward plain) clusters each cross-section separately and links the labels through the mapping step as in Section 2.2. Second, the pooled Ward takes all observations of all units over time and treats them as a single cross-section of $N \times T$ separate units. Third, time-aggregated Ward stacks the $x_{i,t}$ over time into a vector x_t and considers each of its coordinates as one of $N \times D$ separate variables.

		Misclassification		Switching rate	
		Baseline	Half-var.	Baseline	Half-var.
Model	p				
Ward	0.0	0.182	0.101	0.391	0.285
plain	0.01	0.182	0.108	0.392	0.305
	0.1	0.183	0.119	0.408	0.348
	0.25	0.208	0.133	0.431	0.393
Ward	0.0	0.182	0.117	0.368	0.267
pooled	0.01	0.170	0.128	0.368	0.284
	0.1	0.176	0.122	0.386	0.330
	0.25	0.185	0.121	0.423	0.379
Ward	0.0	0.019	0.001		
time-aggregated	0.01	0.045	0.043		
	0.1	0.212	0.195		
	0.25	0.284	0.281		

Table 1: Misclassification rates and switches for the benchmark models. The time-aggregated setting does not allow for switches. The baseline has $\sigma = 1$, and thus large cluster overlaps. For the Half-variance case, $\sigma = 0.5$, and the overlap is smaller.

This last approach does not allow for switches and effectively clusters the whole time series of a unit.

The results are presented in Table 1 and can be compared to the left-hand curves in Figures 5 and B.2. All benchmark approaches produce larger misclassification errors than our new penalized dynamic clustering approach. Only Ward's time-aggregated approach for small p appears to fare slightly better, but at the cost of not allowing for any switches at all. As a consequence, it produces disproportionately large misclassification rates as p increases, exceeding those of the penalized clustering approach of this paper. The difference in misclassification rates are substantial: even for a large set of sub-optimal choices of ε, the new method still beats the benchmarks. For instance, in the baseline design with p set at 0 and 0.01, *any* choice of ε produces lower errors than either the Ward plain or the Ward pooled benchmark. For $p = 0.1$ the misclassification rate in our approach is only higher when $\varepsilon \geq 0.65$. This suggests that also if cluster switches happen more often, a wide range of (optimal or sup-optimal) shrinkage parameters ε results in improvements over the considered benchmarks.

5 Empirical illustration

This section applies the clustering methodology of Section 2 to multivariate panel of $D = 4$ accounting ratios for $N = 28$ European insurance companies' over the period 2010 and 2020 ($T = 11$). Each year, we allocate insurers to one of $k = 1, \ldots, K_t$ distinct business model (peer) groups. We proceed by first describing the data, followed by the empirical results.

5.1 Data

Our sample of $N = 28$ European insurance companies overlaps strongly with a set of 44 insurance companies chosen by EIOPA for its 2021 insurance sector stress test; see EIOPA (2021, Annex A). We observe annual insurer-level accounting data from InsuranceFocus (Bureau van Dijk). We start with the EIOPA's selection of insurers, which together cover approximately 75% of the European Economic Area's insurance market based on total assets. We exclude companies for which a complete set of data is not available, resulting in 26 companies, before adding two large Swiss insurance companies to complement the sample (Swiss Re and Zurich insurance). Table B.1 in Appendix B provides a listing of all firms, along with a subset of estimated cluster allocations.

We select a parsimonious set of variables to classify our selection of European insurers into broadly similar peer-groups. Our choice of variables is motivated by the desire to tell apart four types of insurers: re-insurance, non-life insurance, life insurance, and financial conglomerate. The first three types are insurers that focus on a specific part of the insurance business. The fourth type is a large insurer that owns at least one sizable deposit-taking (bank) subsidiary.

To allocate insurers into peer groups we consider the following variables: insurers' i) ratio of total reinsurance premia received over total traditional (life and non-life) insurance premia; ii) share of life insurance premia to total premia, iii) share of non-life insurance premia to total premia, and iv) share of banking assets (loans and mortgages) to total assets. The first three variables are taken from the insurers' profit-and-loss ("technical

accounts") statements, while the fourth variable is taken from the insurers' consolidated balance sheets. The first variable allows us to distinguish reinsurance firms from "regular" insurers. The second and third variable allow us to further subdivide regular insurers into life- and nonlife insurers. The fourth variable allows us to distinguish financial conglomerates. We rely on International Financial Reporting Standards (IFRS) accounting data, and use domestic-GAAP accounting data when IFRS data are not available.

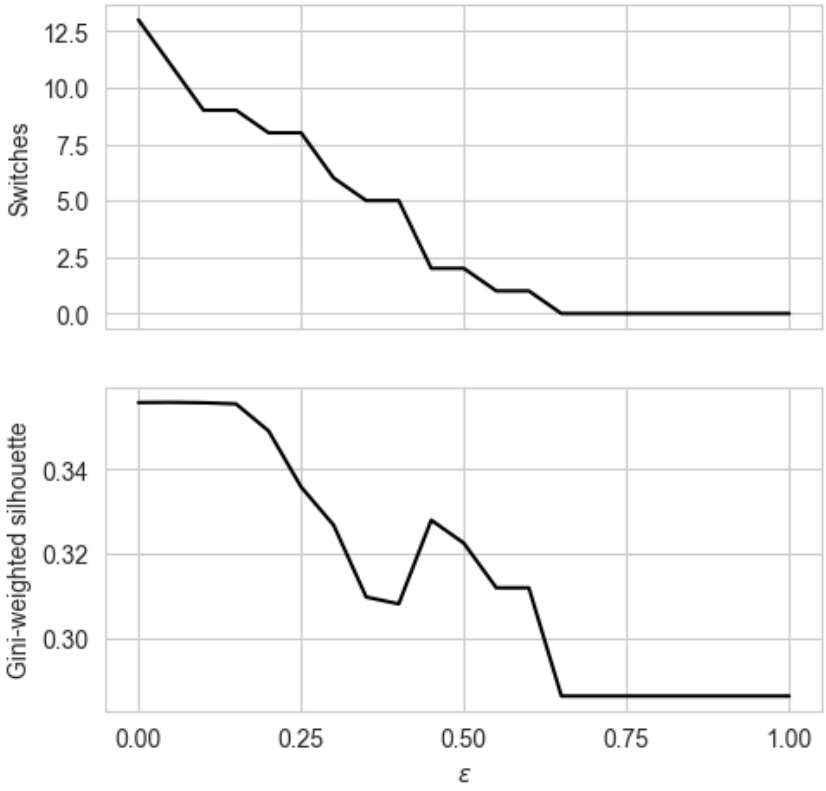
5.2 Clustering outcomes

We first discuss the results for a fixed number of clusters $K = 4$, which is in line with the highest Gini-weighted silhouette index at almost all time points (see below), and our reading of the general industry perception. As a robustness check, we also provide clustering outcomes when K_t is allowed to vary between two and six, corroborating that $K = 4$ is an appropriate choice.

We initialize our clustering method by applying threshold rules to the first cross-section. These threshold rules divide the data into four mutually exclusive and economically interpretable clusters. Firms receiving more reinsurance premia than non-reinsurance premia are allocated to cluster 1 ("reinsurance"). Non-reinsurance firms receiving more than half of their total premium income from life contracts are allocated to cluster 2 ("life"). Non-reinsurance firms receiving most premium income from non-life insurance are allocated to cluster 3 ("non-life"). Firms exhibiting banking assets (total loans and mortgages) of more than a third of total assets are allocated to cluster 4 ("conglomerate"), potentially overriding the other splits. This approach allocates each firm uniquely to one of the four clusters.

We then use our dynamic clustering method, in conjunction with k-means clustering, to allocate the remaining cross-sections conditional on the initial allocation. Our initialization approach has no effect on subsequent cross sections when no shrinkage is imposed ($\varepsilon = 0$). In that case the clustering outcomes quickly revert to the outcomes implied by independent k-means clustering of each cross section in isolation. The higher the amount

Figure 6: Clustering diagnostics as a function of the shrinkage parameter ε. Top panel: total number of cluster switches. Bottom panel: average Gini-weighted silhouette index.



of shrinkage, however, the stickier and the more important the initialization.

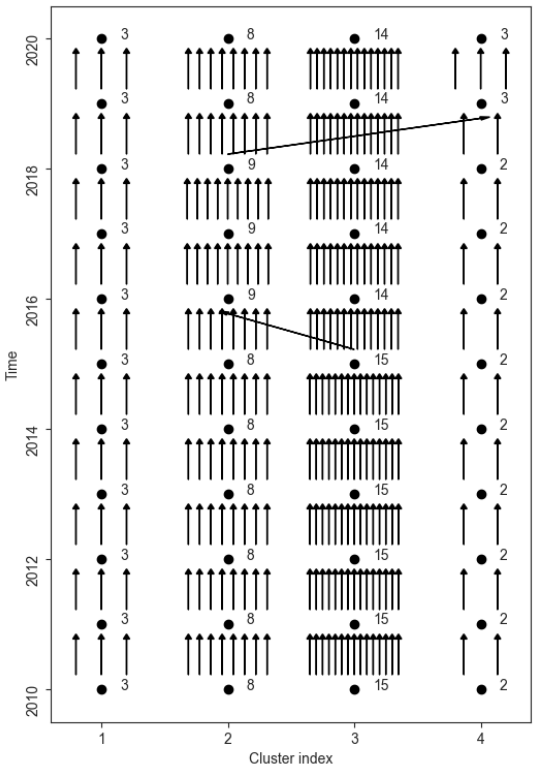
Figure 6 presents clustering diagnostics as a function of the shrinkage parameter ε. Our goal is to decrease the number of incidental switches (flickering) while retaining a high overall fit to the data. Figure 6 allows us to compare the Gini-weighted silhouette index (our measure of fit, in the bottom panel) to the number of switches (in the top panel) associated to each value of ε. The bottom panel of Figure 6 indicates that there is a local maximum in fit at $\varepsilon = 0.45$, coinciding with a low number of cluster switches.[3] After $\varepsilon = 0.45$, the fit decreases sharply. We therefore choose $\varepsilon = 0.45$ for the remainder of the analysis based on $K = 4$.

Our clustering approach leads to stable cluster allocations over time. Figure 7a summarizes our cluster allocation outcomes for $K \equiv 4$ and $\varepsilon = 0.45$. Each column refers to one cluster indexed by $k = 1, \ldots, 4$. Each row denotes one year between 2010 and 2020.

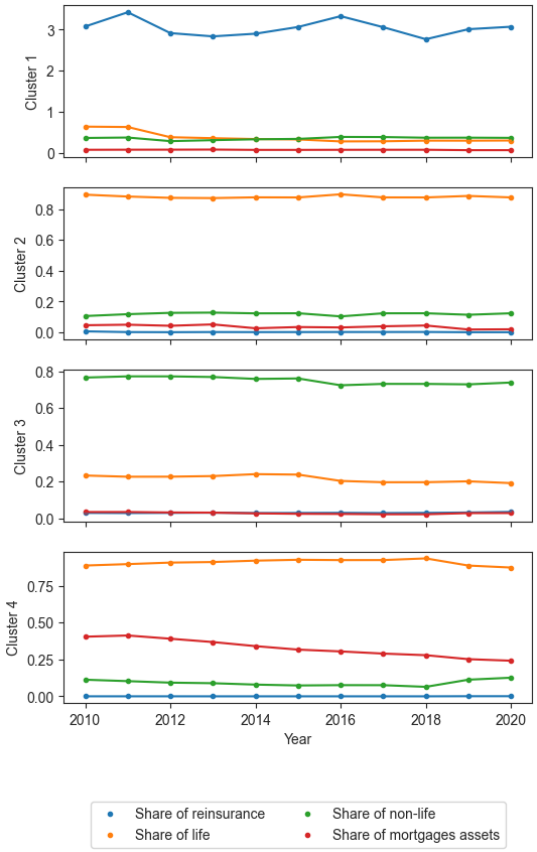
[3]The Gini-weighted silhouette index need not be monotonically decreasing in the shrinkage parameter ε. The clustering outcomes at time t can influence the clustering outcomes at later times, leading to non-monotonicity in the aggregate fit; see also the discussion in Section 2.3.

Figure 7: Clustering composition and transitions and cluster means for $K \equiv 4$ and $\varepsilon = 0.45$. Each column in the left-hand panel refers to one cluster indexed by $k = 1, \ldots, 4$. Each row in the same panel denotes one year between 2010 and 2020, while an arrow represents a transition across clusters. The right hand panel shows the evolution of the cluster means over time.

(a) cluster composition



(b) cluster means

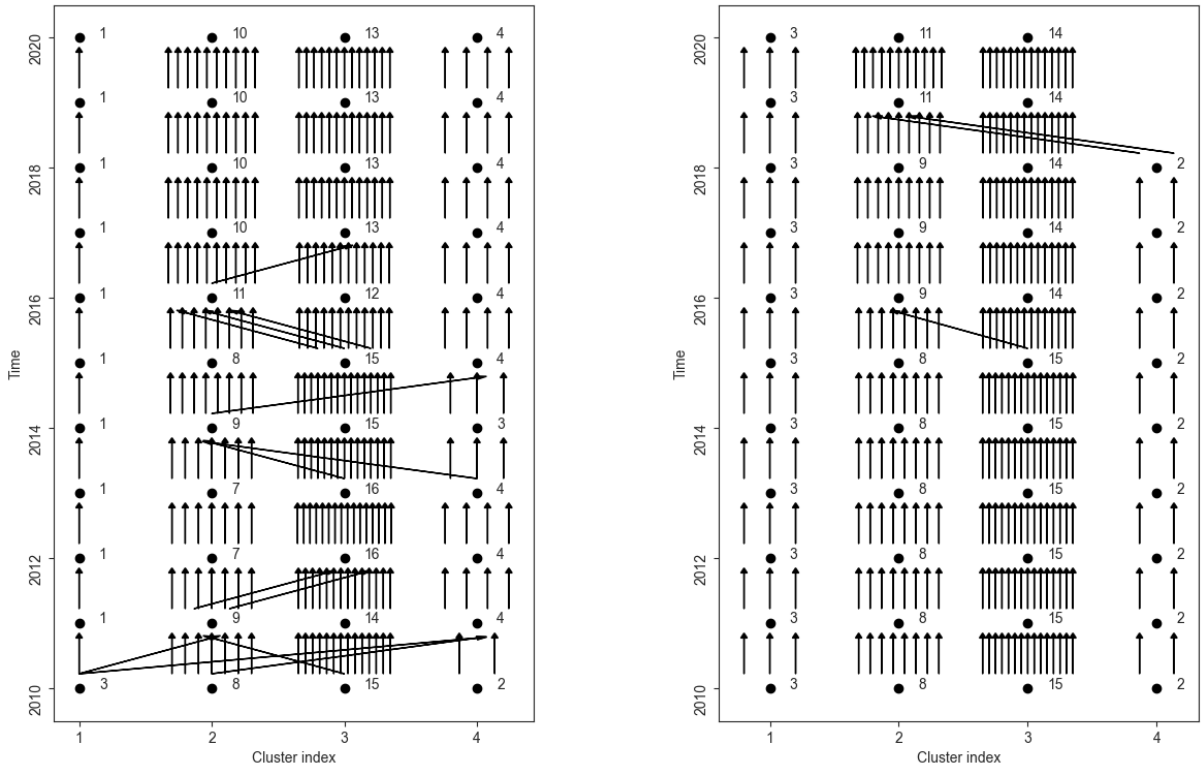


Two cluster transitions are indicated by arrows. The four clusters contain three, eight, fifteen, and two members, respectively, most of the time, with a slight variation in membership only across the last three groups. Traditional non-life and life insurers are the most frequently observed (popular) business models in our sample, ahead of re-insurers and financial conglomerates.

The labels given to each cluster correspond closely with what an inspection of the empirical cluster centroids (means) would suggest. Figure 7b plots the time-varying cluster means for all the variables contained in x_{it}. The first cluster is characterized by a large ratio of reinsurance premia to life and non-life premia. The second and third clusters are characterized by large ratios of life and non-life premia to total non-reinsurance premia,

Figure 8: Clustering composition and transitions for $K \equiv 4$ and $\varepsilon = 0.00$ (panel (a)) as well as $K_t \in \{2, \ldots, 6\}$ and $\varepsilon = 0.55$ (panel (b)). Each column in the figures refers to one cluster. Each row denotes one year between 2010 and 2020. Each arrow represents a transition across clusters.

(a) $\varepsilon = 0.00$; $K \equiv 4$ (b) $\varepsilon = 0.55$; $K_t \in \{2, \ldots, 6\}$



respectively. The fourth cluster is characterized by a substantial ratio of banking assets to total assets.

Figure 8a summarizes the cluster allocation outcomes for $K \equiv 4$ and $\varepsilon = 0$. The clustering outcomes are visibly more volatile, and much harder to interpret economically if no shrinkage is imposed to link the cross-sections over time. Two outcomes are worth noting. First, the reinsurance cluster now shrinks in membership early on in the sample (in 2012), from three to only one member. This can be traced to the first variable being substantially higher for one firm (Swiss Re) than for the other two reinsurance firms (Munich Re, Hannover Re). The fact that the two migrating firms carry "Re" in their names may suggest that these transitions may not necessarily be interpretable. Imposing shrinkage removes these transitions; cf. Figure 7a. Second, there is some noticeable going back- and forth between the life and non-life clusters. This can be traced back to a few

insurers that engage in both life and non-life business, with the precise split between the two being subject to accounting windfalls and other one-off effects (similarly to the setting in Figure 1). Such "middle-of-the-road" or "multi-line" insurers are relatively harder to cluster. Imposing shrinkage avoids these firms flickering back and forth between the life and non-life clusters, yielding more stable clustering outcomes and, in turn, enhancing economic interpretability.

Finally, we allow $K_t \in \{2, 3, 4, 5, 6\}$ to vary over time. As indicated by Algorithm 1, K_t can be chosen to maximize the local time-t silhouette index. We continue to start our clustering algorithm at $K_t = 4$ for $t = 1$, and increase the amount of shrinkage slightly to $\varepsilon = 0.55$ to balance the additional source of instability, trading off goodness-of-fit against clustering stability as before.

Figure 8b presents the clustering outcomes. We note two features. First, four clusters are selected almost always even though K_t is allowed to vary. We see that $K_t = 2$, $K_t = 5$ and $K_t = 6$ are never selected, and that $K_t = 3$ is only selected twice. This supports our initial choice of $K = 4$. Second, the fourth ("conglomerate") cluster appears to merge with the second ("life") cluster late in the sample (in 2019 and 2020). This can be traced back to the two cluster means moving closer together at that time. Whether this corresponds to a permanent "structural" feature of our data going forward is currently unclear and left for future research. Finally, we observe that choosing $\varepsilon > 0$ at a moderately high value allows us to obtain stable clustering results for multivariate panel data both when K_t is time-invariant and when it is time-varying.

6 Conclusion

In this paper, we propose a new approach to clustering in a panel setting, allowing for dynamics in the cluster location, cluster composition and number of clusters, while ensuring stability and persistence of assignments via a shrinkage penalty parameter. The method is widely applicable and extends to many cross-sectional clustering algorithm that produces a distance measure, including, for instance, k-means, k-medians, or hierarchical

clustering. We show how the penalty parameter can be chosen in a data-driven way with a simple weighted version of the well-known silhouette index. We also show analytically and in simulations that selecting a strictly positive shrinkage parameter helps to reduce misclassification in empirically relevant conditions.

An application to business models in the European insurance sector underlines the usefulness of our method in balancing flexibility, i.e. allowing for cluster transitions, and penalizing excessive back-and-forth switching between clusters in economic settings.

References

Ayadi, R., P. Bongini, B. Casu, and D. Cucinelli (2021). Banks' Business Model Migrations in Europe: Determinants and Effects. *British Journal of Management*, 1–20.

Bonhomme, S., T. Lamadon, and E. Manresa (2022). Discretizing unobserved heterogeneity. *Econometrica 90*(2), 625–643.

Bonhomme, S. and E. Manresa (2015). Grouped Patterns of Heterogeneity in Panel Data. *Econometrica 83*(3), 1147–1184.

Catania, L. (2021). Dynamic adaptive mixture models with an application to volatility and risk. *Journal of Financial Econometrics 19*, 531–564.

Cheng, X., F. Schorfheide, and P. Shao (2019). Clustering for multi-dimensional heterogeneity.

Custodio João, I., A. Lucas, J. Schaumburg, and B. Schwaab (2022). Dynamic clustering of multivariate panel data. *Journal of Econometrics*.

David, H. A. (1968). Gini's mean difference rediscovered. *Biometrika 55*(3), 573–575.

EIOPA (2021). 2021 Insurance stress test report. *EIOPA-BoS-21/552 from 16 December 2021, available at www.eiopa.eu* (1), 1–59.

Fox, E. B., E. B. Sudderth, M. I. Jordan, and A. S. Willsky (2011). A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics 5*(2A), 1020–1056.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*, Volume 425. Springer.

Frühwirth-Schnatter, S. (2011). Panel data analysis: a survey on model-based clustering of time series. *Advances in Data Analysis and Classification 5*(4), 251–280.

Frühwirth-Schnatter, S. and G. Malsiner-Walli (2019). From here to infinity: sparse finite versus dirichlet process mixtures in model-based clustering. *Advances in data analysis and classification 13*(1), 33–64.

Fu, W. and P. O. Perry (2020). Estimating the Number of Clusters Using Cross-Validation. *Journal of Computational and Graphical Statistics 29*(1), 162–173.

Grundmann, M., V. Kwatra, M. Han, and I. Essa (2010). Efficient hierarchical graph-based video segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, pp. 2141–2148. IEEE.

Kalnis, P., N. Mamoulis, and S. Bakiras (2005). On Discovering Moving Clusters in Spatio-temporal Data. In C. Bauzer Medeiros, M. J. Egenhofer, and E. Bertino (Eds.), *Advances in Spatial and Temporal Databases*, pp. 364–381. Springer.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly 2*(1-2), 83–97.

Lin, C.-C. and S. Ng (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods 1*(1), 42–55.

Lucas, A., J. Schaumburg, and B. Schwaab (2019). Bank Business Models at Zero Interest Rates. *Journal of Business & Economic Statistics 37*(3), 542–555.

Lumsdaine, R. L., R. Okui, and W. Wang (2022). Estimation of panel group structure models with structural breaks in group memberships and coefficients. *Journal of Econometrics*.

Oliveira, M. and J. Gama (2010). Bipartite Graphs for Monitoring Clusters Transitions. In P. R. Cohen, N. M. Adams, and M. R. Berthold (Eds.), *Advances in Intelligent Data Analysis IX*, pp. 114–124. Springer.

Oliveira, M. and J. Gama (2012). A framework to monitor clusters evolution applied to economy and finance problems. *Intelligent Data Analysis 16*(1), 93–111.

Patton, A. J. and B. M. Weller (2021). Risk Price Variation: The Missing Half of Empirical Asset Pricing. *ERID Working Paper No. 274*.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*, 53–65.

Smith, S. C. (2022). Structural breaks in grouped heterogeneity. *Journal of Business & Economic Statistics*, 1–13.

SSM (2016). SSM SREP methodology booklet. *Available at www.bankingsupervision.europa.eu, accessed on April, 14 2016.*, 1–36.

Wang, Y. and R. S. Tsay (2019). Clustering multiple time series with structural breaks. *Journal of Time Series Analysis 40*(2), 182–202.

Zahn, C. (1971). Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Transactions on Computers C-20*(1), 68–86.

A Proofs of propositions

This appendix presents the proofs of the propositions in Section 3.

Consider a univariate data generating process, where the observation x_t depends on its cluster center c_t, given by

$$x_t = c_t + \eta_t,$$

where η_t has cdf $F(\eta_t)$. The cluster center c_t follows a Markov chain with transition probability p. In this setting, our clustering methodology can be written as

$$\hat{c}_t^{\varepsilon} = \hat{c}^{\varepsilon}(x_t|\hat{c}_{t-1}) = \begin{cases} 1 & \text{if} \quad x_t(1-\varepsilon) + \varepsilon\hat{c}_{t-1} > 1/2, \\ 0 & \text{otherwise.} \end{cases} \tag{A.1}$$

We call this the ε-classifier.

Before discussing the results of Section 3, we present Lemma 1, which will be useful below.

Lemma 1. *Given three real numbers x, a, and b:*

$$|a+x| < |b+x| \iff \begin{cases} x > -(a+b)/2 & \text{if } b > a \\ x < -(a+b)/2 & \text{if } b < a \end{cases}$$

and $|a+x| = |b+x|$ if $a = b$ or $x = -(a+b)/2$.

Proof. Assume $b < a$ and that $|a+x| < |b+x|$. Then $b+x < a+x \leq |a+x| < |b+x| \iff b+x < |b+x|$, which implies that $b + x < 0$. So $|a + x| < -(b + x)$. If $a + x < 0 \iff -a - x > -b - x \iff a < b$, which is a contradiction. So $a + x > 0$. If $a + x > 0 \iff a + x < -b - x \iff x < -(a + b)/2$.

Similarly, assume $b > a$ and that $|a + x| < |b + x|$. Then $a + x < b + x$. If $b + x < 0$ then we would have $|a + x| > |b + x|$, so $b + x > 0$. So $|a + x| < b + x$. If $x \geq -a \iff a + x \geq 0 \implies |a + x| < |b + x|$. If $|a + x| < 0 \iff -a - x < b + x \iff x > -(a + b)/2$. □

In our first result, we define the 1-step-ahead misclassification rate as the misclassification probability at t given the information of the true cluster assignment at $t - 1$. Recall Proposition 1 from Section 3:

Proposition 1. *Given c_{t-1}, the one-step-ahead misclassification probability of the ε-classifier is*

$$\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_t|c_{t-1}) = F\left(\frac{\varepsilon - 1/2}{1-\varepsilon}\right) p + F\left(\frac{-1/2}{1-\varepsilon}\right)(1-p), \tag{9}$$

where F denotes the cdf of η_t.

Proof. First, decompose the misclassification probability in a case where a switch happens at $t-1$, and a case where it does not:

$$\begin{aligned}\mathbb{P}(\hat{c}_t^\varepsilon \neq c_t|c_{t-1}) =&\mathbb{P}(\hat{c}_t^\varepsilon = c_{t-1}|c_t \neq c_{t-1})\mathbb{P}(c_t \neq c_{t-1}) + \mathbb{P}(\hat{c}_t^\varepsilon \neq c_{t-1}|c_t = c_{t-1})\mathbb{P}(c_t = c_{t-1})\\ \mathbb{P}(\hat{c}_t^\varepsilon \neq c_t|c_{t-1}) =&\mathbb{P}(\hat{c}_t^\varepsilon = c_{t-1}|c_t \neq c_{t-1})p + \mathbb{P}(\hat{c}_t^\varepsilon \neq c_{t-1}|c_t = c_{t-1})(1-p) \qquad \text{(A.2)}\end{aligned}$$

(A.2) is composed of two probabilities: the error given a switch (i.e. $\mathbb{P}(\hat{c}_t^\varepsilon = c_{t-1}|c_t \neq c_{t-1})$) and the error given no switch (i.e. $\mathbb{P}(\hat{c}_t^\varepsilon \neq c_{t-1}|c_t = c_{t-1})$). We split the proof in two parts, each calculating one of these two probabilities.

Calculation of the probability of error given a switch $\mathbb{P}(\hat{c}_t^\varepsilon = c_{t-1}|c_t \neq c_{t-1})$**.**

Recalling our definition of $\hat{c}_t^\varepsilon$ in (A.1), we can write the two conditional probabilities in terms of the distance between x_t and the centers, and then in terms of the noise. For $\mathbb{P}(\hat{c}_t^\varepsilon = c_{t-1}|c_t \neq c_{t-1})$ we have

$$\begin{aligned}\mathbb{P}(\hat{c}_t^\varepsilon =&c_{t-1}|c_t \neq c_{t-1})\\ =&\mathbb{P}(|x_t(1-\varepsilon) + \varepsilon c_{t-1} - c_{t-1}| < |x_t(1-\varepsilon) + \varepsilon c_{t-1} - (1-c_{t-1})||c_t \neq c_{t-1})\\ =&\mathbb{P}(|(c_t+\eta_t)(1-\varepsilon) + \varepsilon c_{t-1} - c_{t-1}| < |(c_t+\eta_t)(1-\varepsilon) + \varepsilon c_{t-1} - (1-c_{t-1})||c_t \neq c_{t-1})\\ =&\mathbb{P}(|(1-c_{t-1}+\eta_t)(1-\varepsilon) + \varepsilon c_{t-1} - c_{t-1}| < |(1-c_{t-1}+\eta_t)(1-\varepsilon) + \varepsilon c_{t-1} - 1 + c_{t-1})||c_t \neq c_{t-1})\\ =&\mathbb{P}(|(1-2c_{t-1}+\eta_t)(1-\varepsilon)| < |(1-2c_{t-1}+\eta_t)(1-\varepsilon) - 1 + 2c_{t-1}||c_t \neq c_{t-1})\\ =&\mathbb{P}(|1-2c_{t-1}+\eta_t| < |1-2c_{t-1}+\eta_t + (2c_{t-1}-1)/(1-\varepsilon)||c_t \neq c_{t-1}).\end{aligned}$$

And finally

$$\mathbb{P}(\hat{c}_t^\varepsilon = c_{t-1}|c_t \neq c_{t-1}) = \mathbb{P}(|1-2c_{t-1}+\eta_t| < |1-2c_{t-1}+\eta_t + (2c_{t-1}-1)/(1-\varepsilon)||c_t \neq c_{t-1}). \qquad \text{(A.3)}$$

Applying Lemma 1 we can do away with the absolute value. First, write (A.3) it in terms of a and b:

$$\mathbb{P}(\hat{c}_t^\varepsilon =c_{t-1}|c_t \neq c_{t-1}) = \mathbb{P}(\underbrace{1-2c_{t-1}}_{a}+\eta_t| < |\underbrace{1-2c_{t-1}+(2c_{t-1}-1)/(1-\varepsilon)}_{b}+\eta_t||c_t \neq c_{t-1}).$$

Now check if $a < b$ or $a > b$:

$$\begin{aligned}a < b \iff 1-2c_{t-1} < 1-2c_{t-1}+(2c_{t-1}-1)/(1-\varepsilon) \iff 0 < 2c_{t-1}-1\\ \iff 1/2 < c_{t-1} \iff c_{t-1} = 1.\end{aligned}$$

And the other case:

$$a > b \iff 1 - 2c_{t-1} > 1 - 2c_{t-1} + (2c_{t-1} - 1)/(1-\varepsilon) \iff 1/2 > c_{t-1} \iff c_{t-1} = 0.$$

So we have two cases depending on the true cluster at c_{t-1}. Then, applying Lemma 1 to (A.3) we have

$$\begin{aligned}\mathbb{P}(\hat{c}_t^{\varepsilon} = c_{t-1} | c_t \neq c_{t-1}) &= \mathbb{P}(|1 - 2c_{t-1} + \eta_t| < |1 - 2c_{t-1} + \eta_t + (2c_{t-1} - 1)/(1-\varepsilon)| | c_t \neq c_{t-1}) \\ &= \begin{cases} \mathbb{P}(\eta_t > -\frac{1}{2}(a+b) | c_t \neq c_{t-1}) & \text{if } c_{t-1} = 1 \\ \mathbb{P}(\eta_t < -\frac{1}{2}(a+b) | c_t \neq c_{t-1}) & \text{if } c_{t-1} = 0. \end{cases}\end{aligned}$$

First, calculating the term $-\frac{1}{2}(a+b)$ for each case

$$\begin{aligned}-\tfrac{1}{2}(a+b) &= -\tfrac{1}{2}\left(1 - 2c_{t-1} + 1 - 2c_{t-1} + (2c_{t-1} - 1)/(1-\varepsilon)\right) \\ &= 2c_{t-1} - 1 + \frac{1/2 - c_{t-1}}{1-\varepsilon} \\ -\tfrac{1}{2}(a+b) &= \begin{cases} \dfrac{1/2 - \varepsilon}{1-\varepsilon} & \text{if } c_{t-1} = 1 \\ \dfrac{\varepsilon - 1/2}{1-\varepsilon} & \text{if } c_{t-1} = 0. \end{cases}\end{aligned}$$

Substituting in each of these cases

$$\mathbb{P}(\hat{c}_t^{\varepsilon} = c_{t-1} | c_t \neq c_{t-1}) = \begin{cases} \mathbb{P}(\eta_t > (1/2 - \varepsilon)/(1-\varepsilon) | c_t \neq c_{t-1}) & \text{if } c_{t-1} = 1 \\ \mathbb{P}(\eta_t < (\varepsilon - 1/2)/(1-\varepsilon) | c_t \neq c_{t-1}) & \text{if } c_{t-1} = 0. \end{cases}$$

Using F, the cdf of η_t, and its symmetry we have

$$\begin{aligned}\mathbb{P}(\hat{c}_t^{\varepsilon} = c_{t-1} | c_t \neq c_{t-1}) &= \begin{cases} 1 - F\left((1/2 - \varepsilon)/(1-\varepsilon)\right) & \text{if } c_{t-1} = 0 \\ F\left((\varepsilon - 1/2)/(1-\varepsilon)\right) & \text{if } c_{t-1} = 1 \end{cases} \\ \mathbb{P}(\hat{c}_t^{\varepsilon} = c_{t-1} | c_t \neq c_{t-1}) &= F\left(\frac{\varepsilon - 1/2}{1-\varepsilon}\right). \end{aligned} \tag{A.4}$$

Calculation of the probability of error given no switch $\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_{t-1} | c_t = c_{t-1})$**.**

We follow the same steps as for the probability of error given a switch. We have, for the second

conditional probability on (A.2)

$$
\begin{aligned}
&\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_{t-1} | c_t = c_{t-1}) \\
&\quad = \mathbb{P}(|x_t(1-\varepsilon) + \varepsilon c_{t-1} - c_{t-1}| > |x_t(1-\varepsilon) + \varepsilon c_{t-1} - (1-c_{t-1})| | c_t = c_{t-1}) \\
&\quad = \mathbb{P}(|(c_t+\eta_t)(1-\varepsilon) + \varepsilon c_{t-1} - c_{t-1}| > |(c_t+\eta_t)(1-\varepsilon) + \varepsilon c_{t-1} - (1-c_{t-1})| | c_t = c_{t-1}) \\
&\quad = \mathbb{P}(|\eta_t(1-\varepsilon)| > |\eta_t(1-\varepsilon) - 1 + 2c_{t-1}| | c_t = c_{t-1}).
\end{aligned}
$$

And finally

$$
\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_{t-1} | c_t = c_{t-1}) = \mathbb{P}(|\eta_t| > |\eta_t + (2c_{t-1}-1)/(1-\varepsilon)| | c_t = c_{t-1}). \tag{A.5}
$$

Again we apply Lemma 1 so that we can do away with the absolute value. First, write (A.5) it in terms of a and b:

$$
\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_{t-1} | c_t = c_{t-1}) = \mathbb{P}(\underbrace{0}_{a} + \eta_t| < |\underbrace{(2c_{t-1}-1)/(1-\varepsilon)}_{b} + \eta_t | | c_t = c_{t-1}).
$$

We can immediately check when $a < b$ and $a > b$:

$$
\begin{aligned}
a < b &\iff c_{t-1} = 1 \\
a > b &\iff c_{t-1} = 0,
\end{aligned}
$$

and $-\frac{1}{2}(a+b) = (1/2 - c_{t-1})/(1-\varepsilon)$. Then, applying Lemma 1 to (A.5) we have

$$
\begin{aligned}
\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_{t-1} | c_t = c_{t-1}) &= \mathbb{P}(0 + \eta_t| < |(2c_{t-1}-1)/(1-\varepsilon) + \eta_t | | c_t = c_{t-1}) \\
&= \begin{cases} \mathbb{P}(\eta_t > -\frac{1}{2}(a+b) | c_t = c_{t-1}) & \text{if } \; c_{t-1} = 0 \\ \mathbb{P}(\eta_t < -\frac{1}{2}(a+b) | c_t = c_{t-1}) & \text{if } \; c_{t-1} = 1 \end{cases} \\
&= \begin{cases} \mathbb{P}(\eta_t > +(1/2)/(1-\varepsilon) | c_t = c_{t-1}) & \text{if } \; c_{t-1} = 0 \\ \mathbb{P}(\eta_t < -(1/2)/(1-\varepsilon) | c_t = c_{t-1}) & \text{if } \; c_{t-1} = 1 \end{cases} \\
&= \begin{cases} 1 - F\left((1/2)/(1-\varepsilon)\right) & \text{if } \; c_{t-1} = 0 \\ F\left(-(1/2)/(1-\varepsilon)\right) & \text{if } \; c_{t-1} = 1. \end{cases}
\end{aligned}
$$

Finally, using the symmetry of F,

$$
\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_{t-1} | c_t = c_{t-1}) = F\left(-(1/2)/(1-\varepsilon)\right). \tag{A.6}
$$

We conclude the proof by substituting (A.4) and (A.6) into (A.2), yielding

$$\begin{aligned}\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_t | c_{t-1}) =& \mathbb{P}(\hat{c}_t^{\varepsilon} = c_{t-1} | c_t \neq c_{t-1}) p + \mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_{t-1} | c_t = c_{t-1})(1-p) \\ \mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_t | c_{t-1}) =& F\left(\frac{\varepsilon - 1/2}{1-\varepsilon}\right) p + F\left(-\frac{1/2}{1-\varepsilon}\right)(1-p).\end{aligned}$$

□

Next we prove Proposition 2. Recall that it states:

Proposition 2. *If* $\eta_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$*, the value* ε^* *which minimizes the misclassification rate* (9) *for* $0 < p < \frac{1}{2}$ *is*

$$\varepsilon^* = \frac{2\sigma^2 \log\left(\frac{p}{1-p}\right)}{2\sigma^2 \log\left(\frac{p}{1-p}\right) - 1}. \tag{10}$$

Proof. The proof is a straightforward optimization of the function (9) with $\eta_t \sim N(0, \sigma^2)$. That is,

$$\min_{\varepsilon} \mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_t | c_{t-1}) = \min_{\varepsilon} F\left(\frac{\varepsilon - 1/2}{1-\varepsilon}\right) p + F\left(-\frac{1/2}{1-\varepsilon}\right)(1-p).$$

Taking the derivative we have

$$\frac{\partial \mathbb{P}(.)}{\partial \varepsilon} = p f((\varepsilon - 1/2)/(1-\varepsilon)) \frac{1}{2(1-\varepsilon)^2} + (1-p) f(-1/2(1-\varepsilon))(-1) \frac{1}{2(1-\varepsilon)^2}.$$

Setting it to zero:

$$\begin{aligned}0 &= p f((\varepsilon^* - 1/2)/(1-\varepsilon^*)) \frac{1}{2(1-\varepsilon^*)^2} + (1-p) f(-1/2(1-\varepsilon^*))(-1) \frac{1}{2(1-\varepsilon^*)^2} \\ 0 &= p f((\varepsilon^* - 1/2)/(1-\varepsilon^*)) - (1-p) f(-1/2(1-\varepsilon^*)) \\ 0 &= p \exp\left(-\frac{0.5(\varepsilon^* - 1/2)^2}{(1-\varepsilon^*)^2 \sigma^2}\right) - (1-p) \exp\left(-\frac{0.5}{(2(1-\varepsilon^*))^2 \sigma^2}\right) \\ \frac{0.5}{(2(1-\varepsilon^*))^2 \sigma^2} &= -\log\left(\frac{p}{1-p}\right) + \frac{0.5(\varepsilon^* - 1/2)^2}{(1-\varepsilon^*)^2 \sigma^2} \\ \frac{2^{-2} - (\varepsilon^* - 2^{-1})^2}{(1-\varepsilon^*)^2 \sigma^2} &= -2\log\left(\frac{p}{1-p}\right) \\ \frac{2^{-2} - (\varepsilon^{*2} - \varepsilon^* + 2^{-2})}{(1-\varepsilon^*)^2 \sigma^2} &= -2\log\left(\frac{p}{1-p}\right) \\ \frac{\varepsilon^*}{1-\varepsilon^*} &= -2\sigma^2 \log\left(\frac{p}{1-p}\right) \\ \varepsilon^* &= \frac{2\sigma^2 \log\left(\frac{p}{1-p}\right)}{2\sigma^2 \log\left(\frac{p}{1-p}\right) - 1}.\end{aligned}$$

where f is the Gaussian PDF. □

Corollary 1 extends Proposition 1 to the case of a mean-reverting process.

Corollary 1. *If x_t follows the dynamics $x_t = c_t + \beta(x_{t-1} - c_t) + \eta_t$, then the one-step-ahead probability of error of the ε-classifier is*

$$\begin{aligned}\mathbb{P}(\hat{c}_t^\varepsilon \neq c_t | c_{t-1}) =& F\left(\frac{\varepsilon - 1/2}{1-\varepsilon} + \beta(x_{t-1}(2c_{t-1}-1) + 1 - c_{t-1})\right) p \\ &+ F\left(\frac{-1/2}{1-\varepsilon} + \beta(x_{t-1}(1-2c_{t-1}) + c_{t-1})\right)(1-p),\end{aligned} \tag{11}$$

where F denotes the cdf of η_t.

Proof. This proof follows closely that of Proposition 1.

As before. First decompose the misclassification probability in a case where a switch happens at $t-1$, and a case where it does not:

$$\begin{aligned}\mathbb{P}(\hat{c}_t^\varepsilon \neq c_t | c_{t-1}) =& \mathbb{P}(\hat{c}_t^\varepsilon = c_{t-1} | c_t \neq c_{t-1})\mathbb{P}(c_t \neq c_{t-1}) + \mathbb{P}(\hat{c}_t^\varepsilon \neq c_{t-1} | c_t = c_{t-1})\mathbb{P}(c_t = c_{t-1}) \\ \mathbb{P}(\hat{c}_t^\varepsilon \neq c_t | c_{t-1}) =& \mathbb{P}(\hat{c}_t^\varepsilon = c_{t-1} | c_t \neq c_{t-1})p + \mathbb{P}(\hat{c}_t^\varepsilon \neq c_{t-1} | c_t = c_{t-1})(1-p).\end{aligned} \tag{A.7}$$

We also split the proof in two parts, each calculating one of these two probabilities.

Calculation of the probability of error given a switch $\mathbb{P}(\hat{c}_t^\varepsilon = c_{t-1} | c_t \neq c_{t-1})$**.**

Recalling our definition of $\hat{c}_t^\varepsilon$ in (A.1), we can write the two conditional probabilities in terms of the distance between x_t and the centers, and then in terms of the noise. For $\mathbb{P}(\hat{c}_t^\varepsilon = c_{t-1} | c_t \neq c_{t-1})$,

$$\begin{aligned}\mathbb{P}(\hat{c}_t^\varepsilon =& c_{t-1} | c_t \neq c_{t-1}) \\ =& \mathbb{P}(|x_t(1-\varepsilon) + \varepsilon c_{t-1} - c_{t-1}| < |x_t(1-\varepsilon) + \varepsilon c_{t-1} - (1-c_{t-1})| | c_t \neq c_{t-1}) \\ =& \mathbb{P}(|(c_t + \beta(x_{t-1} - c_t) + \eta_t)(1-\varepsilon) + \varepsilon c_{t-1} - c_{t-1}| \\ & < |(c_t + \beta(x_{t-1} - c_t) + \eta_t)(1-\varepsilon) + \varepsilon c_{t-1} - (1-c_{t-1})| | c_t \neq c_{t-1}) \\ =& \mathbb{P}(|(1 - c_{t-1} + \beta(x_{t-1} - 1 + c_{t-1}) + \eta_t)(1-\varepsilon) + \varepsilon c_{t-1} - c_{t-1}| \\ & < |(1 - c_{t-1} + \beta(x_{t-1} - 1 + c_{t-1}) + \eta_t)(1-\varepsilon) + \varepsilon c_{t-1} - 1 + c_{t-1})| | c_t \neq c_{t-1}).\end{aligned}$$

And finally

$$\begin{aligned}\mathbb{P}(\hat{c}_t^\varepsilon = c_{t-1} | c_t \neq c_{t-1}) =& \mathbb{P}(|1 - 2c_{t-1} + \beta(x_{t-1} - 1 + c_{t-1}) + \eta_t| \\ & < |1 - 2c_{t-1} + \beta(x_{t-1} - 1 + c_{t-1}) + \eta_t + (2c_{t-1} - 1)/(1-\varepsilon)| | c_t \neq c_{t-1}).\end{aligned} \tag{A.8}$$

Applying Lemma 1 we can do away with the absolute value. Writing (A.8) in terms of a and b we have

$$\begin{aligned} a &= 1 - 2c_{t-1} + \beta(x_{t-1} - 1 + c_{t-1}) \\ b &= 1 - 2c_{t-1} + \beta(x_{t-1} - 1 + c_{t-1}) + (2c_{t-1} - 1)/(1 - \varepsilon). \end{aligned}$$

Now check if $a < b$ or $a > b$:

$$\begin{aligned} a < b &\iff 1 - 2c_{t-1} < 1 - 2c_{t-1} + (2c_{t-1} - 1)/(1 - \varepsilon) \iff 0 < 2c_{t-1} - 1 \\ &\iff 1/2 < c_{t-1} \iff c_{t-1} = 1. \end{aligned}$$

And the other case:

$$a > b \iff 1 - 2c_{t-1} > 1 - 2c_{t-1} + (2c_{t-1} - 1)/(1 - \varepsilon) \iff 1/2 > c_{t-1} \iff c_{t-1} = 0$$

So we have two cases depending on the true cluster at c_{t-1}. Applying Lemma 1 to (A.8) we have

$$\begin{aligned} \mathbb{P}(\hat{c}_t^{\varepsilon} = c_{t-1} | c_t \neq c_{t-1}) &= \mathbb{P}(|1 - 2c_{t-1} + \beta(x_{t-1} - 1 + c_{t-1}) + \eta_t| \\ &< |1 - 2c_{t-1} + \beta(x_{t-1} - 1 + c_{t-1}) + \eta_t + (2c_{t-1} - 1)/(1 - \varepsilon)| | c_t \neq c_{t-1}) \\ &= \begin{cases} \mathbb{P}(\eta_t > -\frac{1}{2}(a+b) | c_t \neq c_{t-1}) & \text{if } c_{t-1} = 1 \\ \mathbb{P}(\eta_t < -\frac{1}{2}(a+b) | c_t \neq c_{t-1}) & \text{if } c_{t-1} = 0. \end{cases} \end{aligned}$$

First calculate the term $-\frac{1}{2}(a+b)$ for each case:

$$\begin{aligned} -\tfrac{1}{2}(a+b) &= -\tfrac{1}{2}\left(1 - 2c_{t-1} + 1 - 2c_{t-1} + 2\beta(x_{t-1} - 1 + c_{t-1}) + (2c_{t-1} - 1)/(1 - \varepsilon)\right) \\ &= 2c_{t-1} - 1 + \beta(x_{t-1} - 1 + c_{t-1}) + \frac{1/2 - c_{t-1}}{1 - \varepsilon} \\ -\tfrac{1}{2}(a+b) &= \begin{cases} \beta x_{t-1} + \dfrac{1/2 - \varepsilon}{1 - \varepsilon} & \text{if } c_{t-1} = 1 \\ \beta(x_{t-1} - 1) + \dfrac{\varepsilon - 1/2}{1 - \varepsilon} & \text{if } c_{t-1} = 0. \end{cases} \end{aligned}$$

Substituting in each of these cases we have

$$\mathbb{P}(\hat{c}_t^{\varepsilon} = c_{t-1} | c_t \neq c_{t-1}) = \begin{cases} \mathbb{P}(\eta_t > \beta x_{t-1} + (1/2 - \varepsilon)/(1 - \varepsilon) | c_t \neq c_{t-1}) & \text{if } c_{t-1} = 1 \\ \mathbb{P}(\eta_t < \beta(x_{t-1} - 1) + (\varepsilon - 1/2)/(1 - \varepsilon) | c_t \neq c_{t-1}) & \text{if } c_{t-1} = 0. \end{cases}$$

Using F, the cdf of η_t, and its symmetry,

$$\mathbb{P}(\hat{c}_t^{\varepsilon} = c_{t-1} | c_t \neq c_{t-1}) = \begin{cases} 1 - F\left(\beta x_{t-1} + (1/2 - \varepsilon)/(1-\varepsilon)\right) & \text{if } c_{t-1} = 0 \\ F\left(\beta(x_{t-1} - 1) + (\varepsilon - 1/2)/(1-\varepsilon)\right) & \text{if } c_{t-1} = 1. \end{cases}$$

Which we can write more compactly as

$$\mathbb{P}(\hat{c}_t^{\varepsilon} = c_{t-1} | c_t \neq c_{t-1}) = F\left(\beta(x_{t-1}(2c_{t-1} - 1) + 1 - c_{t-1}) + \frac{\varepsilon - 1/2}{1-\varepsilon}\right). \tag{A.9}$$

Calculation of the probability of error given no switch $\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_{t-1} | c_t = c_{t-1})$**.**

We follow the same steps as for the probability of error given a switch. We have, for the second conditional probability on (A.2),

$$\begin{aligned}
&\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_{t-1} | c_t = c_{t-1}) \\
&\quad = \mathbb{P}(|x_t(1-\varepsilon) + \varepsilon c_{t-1} - c_{t-1}| > |x_t(1-\varepsilon) + \varepsilon c_{t-1} - (1 - c_{t-1})| | c_t = c_{t-1}) \\
&\quad = \mathbb{P}(|(c_t + \beta(x_{t-1} - c_t) + \eta_t)(1-\varepsilon) + \varepsilon c_{t-1} - c_{t-1}| \\
&\qquad > |(c_t + \beta(x_{t-1} - c_t) + \eta_t)(1-\varepsilon) + \varepsilon c_{t-1} - (1 - c_{t-1})| | c_t = c_{t-1}) \\
&\quad = \mathbb{P}(|(\beta(x_{t-1} - c_{t-1}) + \eta_t)(1-\varepsilon)| > |(\beta(x_{t-1} - c_{t-1}) + \eta_t)(1-\varepsilon) - 1 + 2c_{t-1}| | c_t = c_{t-1}).
\end{aligned}$$

And finally

$$\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_{t-1} | c_t = c_{t-1}) = \mathbb{P}(|\beta(x_{t-1} - c_{t-1}) + \eta_t| > |\beta(x_{t-1} - c_{t-1}) + \eta_t + (2c_{t-1} - 1)/(1-\varepsilon)| | c_t = c_{t-1}).$$

Again we apply Lemma 1 so that we can do away with the absolute value. Writing the equation above in terms of a and b we have:

$$\begin{aligned}
a &= \beta(x_{t-1} - c_{t-1}) \\
b &= \beta(x_{t-1} - c_{t-1}) + (2c_{t-1} - 1)/(1-\varepsilon).
\end{aligned}$$

Now check if $a < b$ or $a > b$:

$$\begin{aligned}
a < b &\iff c_{t-1} = 1 \\
a > b &\iff c_{t-1} = 0,
\end{aligned}$$

and $-\frac{1}{2}(a+b) = -\beta(x_{t-1} - c_{t-1}) + (1/2 - c_{t-1})/(1-\varepsilon)$. Then, applying Lemma 1 we have

$$\begin{aligned}
\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_{t-1} | c_t = c_{t-1}) &= \mathbb{P}(\beta(x_{t-1} - c_{t-1}) + \eta_t| < |\beta(x_{t-1} - c_{t-1}) + (2c_{t-1} - 1)/(1-\varepsilon) + \eta_t || c_t = c_{t-1}) \\
&= \begin{cases} \mathbb{P}(\eta_t > -\frac{1}{2}(a+b) | c_t = c_{t-1}) & \text{if } c_{t-1} = 0 \\ \mathbb{P}(\eta_t < -\frac{1}{2}(a+b) | c_t = c_{t-1}) & \text{if } c_{t-1} = 1 \end{cases} \\
&= \begin{cases} \mathbb{P}(\eta_t > -\beta x_{t-1} + (1/2)/(1-\varepsilon) | c_t = c_{t-1}) & \text{if } c_{t-1} = 0 \\ \mathbb{P}(\eta_t < -\beta(x_{t-1} - 1) - (1/2)/(1-\varepsilon) | c_t = c_{t-1}) & \text{if } c_{t-1} = 1 \end{cases} \\
&= \begin{cases} 1 - F\left(-\beta x_{t-1} + (1/2)/(1-\varepsilon)\right) & \text{if } c_{t-1} = 0 \\ F\left(-\beta(x_{t-1} - 1) - (1/2)/(1-\varepsilon)\right) & \text{if } c_{t-1} = 1. \end{cases}
\end{aligned}$$

Finally, using the symmetry of F,

$$\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_{t-1} | c_t = c_{t-1}) = F\left(\beta(x_{t-1}(1 - 2c_{t-1}) + c_{t-1}) - \frac{1/2}{1-\varepsilon}\right). \tag{A.10}$$

We conclude the proof by substituting (A.9) and (A.10) into (A.7), yielding

$$\begin{aligned}
\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_t | c_{t-1}) =& \mathbb{P}(\hat{c}_t^{\varepsilon} = c_{t-1} | c_t \neq c_{t-1}) p + \mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_{t-1} | c_t = c_{t-1})(1-p) \\
\mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_t | c_{t-1}) =& F\left(\frac{\varepsilon - 1/2}{1-\varepsilon} + \beta(x_{t-1}(2c_{t-1} - 1) + 1 - c_{t-1})\right) p \\
&+ F\left(\frac{-1/2}{1-\varepsilon} + \beta(x_{t-1}(1 - 2c_{t-1}) + c_{t-1})\right)(1-p).
\end{aligned}$$

□

Proposition 3 states the probabilities of correct classification in a recursive form.

Proposition 3. *The conditional correct classification probabilities q_t follow the recursion*

$$q_{t+1} = \begin{pmatrix} \frac{z_{00} \cdot (1-p) \cdot \pi_{0,t}}{\pi_{0,t+1}} - \frac{z_{10} \cdot (1-p) \cdot \pi_{0,t}}{\pi_{0,t+1}} & \frac{z_{10} \cdot p \cdot \pi_{1,t}}{\pi_{0,t+1}} - \frac{z_{00} \cdot p \cdot \pi_{1,t}}{\pi_{0,t+1}} \\ \frac{z_{01} \cdot p \cdot \pi_{0,t}}{\pi_{1,t+1}} - \frac{z_{11} \cdot p \cdot \pi_{0,t}}{\pi_{1,t+1}} & \frac{z_{11} \cdot (1-p) \cdot \pi_{1,t}}{\pi_{1,t+1}} - \frac{z_{01} \cdot (1-p) \cdot \pi_{1,t}}{\pi_{1,t+1}} \end{pmatrix} \cdot q_t \tag{12}$$

$$+ \begin{pmatrix} \frac{z_{00} \cdot p \cdot \pi_{1,t}}{\pi_{0,t+1}} + \frac{z_{10} \cdot (1-p) \cdot \pi_{0,t}}{\pi_{0,t+1}} \\ \frac{z_{01} \cdot (1-p) \cdot \pi_{1,t}}{\pi_{1,t+1}} + \frac{z_{11} \cdot p \cdot \pi_{0,t}}{\pi_{1,t+1}}, \end{pmatrix}$$

where

$$z_{i0} = F\left(\frac{1/2 - i \cdot \varepsilon}{1-\varepsilon}\right), \quad z_{i1} = 1 - F\left(\frac{1/2 - i \cdot \varepsilon}{1-\varepsilon} - 1\right).$$

Proof. Define the marginal probability for the true state as $\pi_{i,t} = \mathbb{P}(c_t = i)$. Also define the conditional

correct classification probabilities

$$q_t = \begin{pmatrix} q_{0,t} \\ q_{1,t} \end{pmatrix} = \begin{pmatrix} \mathbb{P}(\hat{c}^{\epsilon}_{t+1} = 0 \mid c_{t+1} = 0) \\ \mathbb{P}(\hat{c}^{\epsilon}_{t+1} = 1 \mid c_{t+1} = 1) \end{pmatrix}.$$

The correct classification probability is now given by

$$q_{0,t}\pi_{0,t} + q_{1,t}\pi_{1,t}.$$

Finally, define

$$\begin{aligned} z_{00} &= F\left(\frac{0.5 - \epsilon \cdot 0}{1 - \epsilon} - 0\right), \\ z_{10} &= F\left(\frac{0.5 - \epsilon \cdot 1}{1 - \epsilon} - 0\right), \\ z_{01} &= 1 - F\left(\frac{0.5 - \epsilon \cdot 0}{1 - \epsilon} - 1\right), \\ z_{11} &= 1 - F\left(\frac{0.5 - \epsilon \cdot 1}{1 - \epsilon} - 1\right). \end{aligned}$$

We split the proof into the calculation of the terms $q_{0,t+1}$ and $q_{1,t+1}$.

Note that

$$q_{0,t+1} = \mathbb{P}(\hat{c}^{\epsilon}_{t+1} = 0 \mid c_{t+1} = 0)$$

$$= \mathbb{P}(\hat{c}^{\epsilon}_{t+1} = 0 \mid \hat{c}^{\epsilon}_{t} = 0, c_{t+1} = 0) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0 \mid c_{t+1} = 0) +$$
$$\mathbb{P}(\hat{c}^{\epsilon}_{t+1} = 0 \mid \hat{c}^{\epsilon}_{t} = 1, c_{t+1} = 0) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1 \mid c_{t+1} = 0)$$

$$= F\left(\frac{0.5 - \epsilon \cdot 0}{1 - \epsilon} - 0\right) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0 \mid c_{t+1} = 0) +$$
$$F\left(\frac{0.5 - \epsilon \cdot 1}{1 - \epsilon} - 0\right) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1 \mid c_{t+1} = 0)$$

$$= z_{00} \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0 \mid c_{t+1} = 0) + z_{10} \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1 \mid c_{t+1} = 0)$$

$$= z_{00} \cdot \frac{\mathbb{P}(\hat{c}^{\epsilon}_{t} = 0, c_{t+1} = 0, c_t = 0) + \mathbb{P}\hat{c}^{\epsilon}_{t} = 0, c_{t+1} = 0, c_t = 1)}{\mathbb{P}(c_{t+1} = 0)} +$$
$$z_{10} \cdot \frac{\mathbb{P}(\hat{c}^{\epsilon}_{t} = 1, c_{t+1} = 0, c_t = 0) + \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1, c_{t+1} = 0, c_t = 1)}{\mathbb{P}(c_{t+1} = 0)}$$

$$= \frac{z_{00}}{\pi_{0,t+1}} \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0, c_{t+1} = 0, c_t = 0) +$$
$$\frac{z_{00}}{\pi_{0,t+1}} \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0, c_{t+1} = 0, c_t = 1) +$$
$$\frac{z_{10}}{\pi_{0,t+1}} \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1, c_{t+1} = 0, c_t = 0) +$$
$$\frac{z_{10}}{\pi_{0,t+1}} \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1, c_{t+1} = 0, c_t = 1)$$

$$= \frac{z_{00}}{\pi_{0,t+1}} \cdot \mathbb{P}(c_{t+1} = 0 \mid c_t = 0) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0 \mid c_t = 0) \cdot \mathbb{P}(c_t = 0) +$$
$$\frac{z_{00}}{\pi_{0,t+1}} \cdot \mathbb{P}(c_{t+1} = 0 \mid c_t = 1) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0 \mid c_t = 1) \cdot \mathbb{P}(c_t = 1) +$$
$$\frac{z_{10}}{\pi_{0,t+1}} \cdot \mathbb{P}(c_{t+1} = 0 \mid c_t = 0) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1 \mid c_t = 0) \cdot \mathbb{P}(c_t = 0) +$$
$$\frac{z_{10}}{\pi_{0,t+1}} \cdot \mathbb{P}(c_{t+1} = 0 \mid c_t = 1) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1 \mid c_t = 1) \cdot \mathbb{P}(c_t = 1)$$

$$= \frac{z_{00}}{\pi_{0,t+1}} \cdot (1 - p) \cdot q_{0,t} \cdot \pi_{0,t} + \frac{z_{00}}{\pi_{0,t+1}} \cdot p \cdot (1 - q_{1,t}) \cdot \pi_{1,t} +$$
$$\frac{z_{10}}{\pi_{0,t+1}} \cdot (1 - p) \cdot (1 - q_{0,t}) \cdot \pi_{0,t} + \frac{z_{10}}{\pi_{0,t+1}} \cdot p \cdot q_{1,t} \cdot \pi_{1,t}.$$

We also have

$$q_{1,t+1} = \mathbb{P}(\hat{c}^{\epsilon}_{t+1} = 1 \mid c_{t+1} = 1)$$

$$= \mathbb{P}(\hat{c}^{\epsilon}_{t+1} = 1 \mid \hat{c}^{\epsilon}_{t} = 0, c_{t+1} = 1) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0 \mid c_{t+1} = 1) +$$
$$\mathbb{P}(\hat{c}^{\epsilon}_{t+1} = 1 \mid \hat{c}^{\epsilon}_{t} = 1, c_{t+1} = 1) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1 \mid c_{t+1} = 1)$$

$$= \left(1 - F\left(\frac{0.5 - \epsilon \cdot 0}{1 - \epsilon} - 1\right)\right) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0 \mid c_{t+1} = 1) +$$
$$\left(1 - F\left(\frac{0.5 - \epsilon \cdot 1}{1 - \epsilon} - 1\right)\right) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1 \mid c_{t+1} = 1)$$

$$= z_{01} \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0 \mid c_{t+1} = 1) + z_{11} \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1 \mid c_{t+1} = 1)$$

$$= z_{01} \cdot \frac{\mathbb{P}(\hat{c}^{\epsilon}_{t} = 0, c_{t+1} = 0, c_t = 0) + \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0, c_{t+1} = 0, c_t = 1)}{\mathbb{P}(c_{t+1} = 0)} +$$
$$z_{11} \cdot \frac{\mathbb{P}(\hat{c}^{\epsilon}_{t} = 1, c_{t+1} = 0, c_t = 0) + \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1, c_{t+1} = 0, c_t = 1)}{\mathbb{P}(c_{t+1} = 0)}$$

$$= \frac{z_{01}}{\pi_{1,t+1}} \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0, c_{t+1} = 1, c_t = 0) +$$
$$\frac{z_{01}}{\pi_{1,t+1}} \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0, c_{t+1} = 1, c_t = 1) +$$
$$\frac{z_{11}}{\pi_{1,t+1}} \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1, c_{t+1} = 1, c_t = 0) +$$
$$\frac{z_{11}}{\pi_{1,t+1}} \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1, c_{t+1} = 1, c_t = 1)$$

$$= \frac{z_{01}}{\pi_{1,t+1}} \cdot \mathbb{P}(c_{t+1} = 1 \mid c_t = 0) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0 \mid c_t = 0) \cdot \mathbb{P}(c_t = 0) +$$
$$\frac{z_{01}}{\pi_{1,t+1}} \cdot \mathbb{P}(c_{t+1} = 1 \mid c_t = 1) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 0 \mid c_t = 1) \cdot \mathbb{P}(c_t = 1) +$$
$$\frac{z_{11}}{\pi_{1,t+1}} \cdot \mathbb{P}(c_{t+1} = 1 \mid c_t = 0) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1 \mid c_t = 0) \cdot \mathbb{P}(c_t = 0) +$$
$$\frac{z_{11}}{\pi_{1,t+1}} \cdot \mathbb{P}(c_{t+1} = 1 \mid c_t = 1) \cdot \mathbb{P}(\hat{c}^{\epsilon}_{t} = 1 \mid c_t = 1) \cdot \mathbb{P}(c_t = 1)$$

$$= \frac{z_{01}}{\pi_{1,t+1}} \cdot p \cdot q_{0,t} \cdot \pi_{0,t} + \frac{z_{01}}{\pi_{1,t+1}} \cdot (1-p) \cdot (1 - q_{1,t}) \cdot \pi_{1,t} +$$
$$\frac{z_{11}}{\pi_{1,t+1}} \cdot p \cdot (1 - q_{0,t}) \cdot \pi_{0,t} + \frac{z_{11}}{\pi_{1,t+1}} \cdot (1-p) \cdot q_{1,t} \cdot \pi_{1,t}.$$

Putting $q_{0,t+1}$ and $q_{1,t+1}$ together in a system of equations, we can write

$$q_{t+1} = \begin{pmatrix} \frac{z_{00}\cdot(1-p)\cdot\pi_{0,t}}{\pi_{0,t+1}} - \frac{z_{10}\cdot(1-p)\cdot\pi_{0,t}}{\pi_{0,t+1}} & \frac{z_{10}\cdot p\cdot\pi_{1,t}}{\pi_{0,t+1}} - \frac{z_{00}\cdot p\cdot\pi_{1,t}}{\pi_{0,t+1}} \\ \frac{z_{01}\cdot p\cdot\pi_{0,t}}{\pi_{1,t+1}} - \frac{z_{11}\cdot p\cdot\pi_{0,t}}{\pi_{1,t+1}} & \frac{z_{11}\cdot(1-p)\cdot\pi_{1,t}}{\pi_{1,t+1}} - \frac{z_{01}\cdot(1-p)\cdot\pi_{1,t}}{\pi_{1,t+1}} \end{pmatrix} \cdot q_t$$

$$+ \begin{pmatrix} \frac{z_{00}\cdot p\cdot\pi_{1,t}}{\pi_{0,t+1}} + \frac{z_{10}\cdot(1-p)\cdot\pi_{0,t}}{\pi_{0,t+1}} \\ \frac{z_{01}\cdot(1-p)\cdot\pi_{1,t}}{\pi_{1,t+1}} + \frac{z_{11}\cdot p\cdot\pi_{0,t}}{\pi_{1,t+1}} \end{pmatrix}.$$

□

Corollary 2. *The limiting probabilities of correct classification q for a symmetric Markov chain $\mathbb{P}(c_t = 1|c_{t-1} = 0) = \mathbb{P}(c_t = 0|c_{t-1} = 1) = p$ are*

$$q = \begin{pmatrix} 1-(1-p)(z_{00}-z_{10}) & p\cdot(z_{00}-z_{10}) \\ p\cdot(z_{11}-z_{01}) & 1-(1-p)(z_{11}-z_{01}) \end{pmatrix}^{-1} \times \begin{pmatrix} z_{10}+p\cdot(z_{00}-z_{10}) \\ z_{01}+p\cdot(z_{11}-z_{01}) \end{pmatrix}.$$

The corresponding limiting misclassification probability is

$$\lim_{t\to\infty} \mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_t) = 1 - \frac{1}{2}\frac{z_{01}(1-\tilde{z}_{00}) + z_{10}(1-\tilde{z}_{11}) + p(\tilde{z}_{00}+\tilde{z}_{11}-2\tilde{z}_{11}\tilde{z}_{00})}{1-(1-p)(\tilde{z}_{00}+\tilde{z}_{11}) + (1-2p)\tilde{z}_{00}\tilde{z}_{11}}, \tag{13}$$

where $\tilde{z}_{00} = z_{00} - z_{10}$ and $\tilde{z}_{11} = z_{11} - z_{01}$.

Proof. First note that for the current symmetric Markov chain $\lim_{t\to\infty} \pi_{i,t} = 0.5$. Then, the statement in Proposition 3 becomes:

$$\begin{aligned} q_{t+1} &= \begin{pmatrix} z_{00}\cdot(1-p) - z_{10}\cdot(1-p) & z_{10}\cdot p - z_{00}\cdot p \\ z_{01}\cdot p - z_{11}\cdot p & z_{11}\cdot(1-p) - z_{01}\cdot(1-p) \end{pmatrix} \cdot q_t \\ &\quad + \begin{pmatrix} z_{00}\cdot p + z_{10}\cdot(1-p) \\ z_{01}\cdot(1-p) + z_{11}\cdot p \end{pmatrix} \\ q_{t+1} &= A\cdot q_t + b. \end{aligned}$$

At the limit, $q_{t+1} = q_t = q$ and so

$$(I_2 - A)q = b.$$

Using this and writing $\tilde{z}_{00} = z_{00} - z_{10}$ and $\tilde{z}_{11} = z_{11} - z_{01}$, we can solve for

$$q = \begin{pmatrix} 1-(1-p)\tilde{z}_{00} & p\cdot\tilde{z}_{00} \\ p\cdot\tilde{z}_{11} & 1-(1-p)\tilde{z}_{11} \end{pmatrix}^{-1} \times \begin{pmatrix} z_{10}+p\cdot\tilde{z}_{00} \\ z_{01}+p\cdot\tilde{z}_{11} \end{pmatrix}.$$

Calculating the inverse we get

$$(I_2 - A)^{-1} = \begin{pmatrix} \frac{-p\tilde{z}_{11}+\tilde{z}_{11}-1}{p\tilde{z}_{00}(2\tilde{z}_{11}-1)-p\tilde{z}_{11}+\tilde{z}_{00}(-\tilde{z}_{11})+\tilde{z}_{00}+\tilde{z}_{11}-1} & \frac{p\tilde{z}_{00}}{p\tilde{z}_{00}(2\tilde{z}_{11}-1)-p\tilde{z}_{11}+\tilde{z}_{00}(-\tilde{z}_{11})+\tilde{z}_{00}+\tilde{z}_{11}-1} \\ \frac{p\tilde{z}_{11}}{p\tilde{z}_{00}(2\tilde{z}_{11}-1)-p\tilde{z}_{11}+\tilde{z}_{00}(-\tilde{z}_{11})+\tilde{z}_{00}+\tilde{z}_{11}-1} & \frac{-p\tilde{z}_{00}+\tilde{z}_{00}-1}{p\tilde{z}_{00}(2\tilde{z}_{11}-1)-p\tilde{z}_{11}+\tilde{z}_{00}(-\tilde{z}_{11})+\tilde{z}_{00}+\tilde{z}_{11}-1} \end{pmatrix},$$

and,

$$q = (I_2 - A)^{-1}b = \begin{pmatrix} \frac{p\tilde{z}_{00}(z_{01}+\tilde{z}_{11}-1)-pz_{10}\tilde{z}_{11}+z_{10}(\tilde{z}_{11}-1)}{p\tilde{z}_{00}(2\tilde{z}_{11}-1)-p\tilde{z}_{11}+\tilde{z}_{00}(-\tilde{z}_{11})+\tilde{z}_{00}+\tilde{z}_{11}-1} \\ \frac{z_{01}(-p\tilde{z}_{00}+\tilde{z}_{00}-1)+p\tilde{z}_{11}(\tilde{z}_{00}+z_{10}-1)}{p\tilde{z}_{00}(2\tilde{z}_{11}-1)-p\tilde{z}_{11}+\tilde{z}_{00}(-\tilde{z}_{11})+\tilde{z}_{00}+\tilde{z}_{11}-1} \end{pmatrix}.$$

The limiting misclassification probability $(0.5, 0.5) \cdot q$ then is

$$\lim_{t\to\infty} \mathbb{P}(\hat{c}_t^{\varepsilon} \neq c_t) = 1 - \frac{1}{2}\frac{z_{01}(1-\tilde{z}_{00}) + z_{10}(1-\tilde{z}_{11}) + p(\tilde{z}_{00}+\tilde{z}_{11}-2\tilde{z}_{11}\tilde{z}_{00})}{1-(1-p)(\tilde{z}_{00}+\tilde{z}_{11})+(1-2p)\tilde{z}_{00}\tilde{z}_{11}}.$$

□

Corollary 3. *Let $f(\eta_t)$ be the pdf of η_t, corresponding to the cdf $F(\eta_t)$. Then under the same conditions as Corollary 2, the derivative of the limiting misclassification probability at $\varepsilon = 0$ is given by*

$$\frac{\partial\ \frac{1}{2}(1-(1\ ,\ 1)\ q)}{\partial\varepsilon}\bigg|_{\varepsilon=0} = \tfrac{1}{4}\big(f(\tfrac{1}{2}) - f(-\tfrac{1}{2})\big) + \tfrac{1}{2}p\left(f(-\tfrac{1}{2})F(\tfrac{1}{2}) - f(\tfrac{1}{2})F(-\tfrac{1}{2})\right) + \tfrac{1}{2}(p-1)\left(f(\tfrac{1}{2})F(\tfrac{1}{2}) - f(-\tfrac{1}{2})F(-\tfrac{1}{2})\right).$$

If the pdf f is symmetric around zero, this expression simplifies to

$$\frac{\partial\ \frac{1}{2}(1-(1\ ,\ 1)\ q)}{\partial\varepsilon}\bigg|_{\varepsilon=0} = -\tfrac{1}{2}(1-2p)\ f(\tfrac{1}{2})\left(2F(\tfrac{1}{2})-1\right),$$

which is negative for $p < 0.5$.

Proof. From Corollary 2, write the limiting misclassification probability

$$1 - q = 1 - \tfrac{1}{2}(1,1)A^{-1}b \tag{A.11}$$

with

$$A = \begin{pmatrix} 1-(1-p)(z_{00}-z_{10}) & p\cdot(z_{00}-z_{10}) \\ p\cdot(z_{11}-z_{01}) & 1-(1-p)(z_{11}-z_{01}) \end{pmatrix}$$

and

$$b = \begin{pmatrix} z_{10} + p\cdot(z_{00}-z_{10}) \\ z_{01} + p\cdot(z_{11}-z_{01}) \end{pmatrix}.$$

Let also ∇z denote $\partial z/\partial\varepsilon\,|_{\varepsilon=0}$. The derivative of (A.11) can be written

$$-\tfrac{1}{2}|A|^{-2}\left(|A|\cdot(1\ ,\ 1)(\nabla A^{\star})b+|A|\cdot(1\ ,\ 1)A^{\star}(\nabla b)-(\nabla|A|)\cdot(1\ ,\ 1)A^{\star}b\right), \tag{A.12}$$

where $A^{\star}$ denotes the transposed matrix of co-factors such that $A^{-1}=A^{\star}/|A|$. Define $f(\eta)=\mathrm{d}F(\eta)\,/\,\mathrm{d}\eta$. Then:

$$\begin{aligned}
z_{00} &= F\left(\frac{0.5-\epsilon\cdot 0}{1-\epsilon}-0\right) \implies \nabla z_{00}=\tfrac{1}{2}f(\tfrac{1}{2}),\\
z_{10} &= F\left(\frac{0.5-\epsilon\cdot 1}{1-\epsilon}-0\right) \implies \nabla z_{10}=-\tfrac{1}{2}f(\tfrac{1}{2}),\\
z_{01} &= 1-F\left(\frac{0.5-\epsilon\cdot 0}{1-\epsilon}-1\right) \implies \nabla z_{01}=-\tfrac{1}{2}f(-\tfrac{1}{2}),\\
z_{11} &= 1-F\left(\frac{0.5-\epsilon\cdot 1}{1-\epsilon}-1\right) \implies \nabla z_{11}=\tfrac{1}{2}f(-\tfrac{1}{2}),\\
z_{00}\big|_{\varepsilon=0} &= z_{10}\big|_{\varepsilon=0}=F(\tfrac{1}{2}),\\
z_{01}\big|_{\varepsilon=0} &= z_{11}\big|_{\varepsilon=0}=1-F(-\tfrac{1}{2}).
\end{aligned}$$

And also, for A and b,

$$\begin{aligned}
A\big|_{\varepsilon=0} &= A^{\star}\big|_{\varepsilon=0}=I_2,\\
|A|\ \big|_{\varepsilon=0} &= 1,\\
\nabla A^{\star} &= -\begin{pmatrix}(1-p)\ f(-\frac{1}{2}) & p\ f(\frac{1}{2})\\ p\ f(-\frac{1}{2}) & (1-p)\ f(\frac{1}{2})\end{pmatrix},\\
\nabla|A| &= -(1-p)f(\tfrac{1}{2})-(1-p)f(-\tfrac{1}{2})\\
&= -(1-p)\big(f(\tfrac{1}{2})+f(-\tfrac{1}{2})\big),\\
b\big|_{\varepsilon=0} &= \big(F(\tfrac{1}{2})\ ,\ 1-F(-\tfrac{1}{2})\big)',\\
\nabla b &= \big(-\tfrac{1}{2}f(\tfrac{1}{2})+p\ f(\tfrac{1}{2})\ ,\ -\tfrac{1}{2}f(-\tfrac{1}{2})+p\ f(-\tfrac{1}{2})\big)'\\
&= -\tfrac{1}{2}\left(f(\tfrac{1}{2})(1-2p)\ ,\ f(-\tfrac{1}{2})(1-2p)\right)'\\
&= -\tfrac{1}{2}(1-2p)\left(f(\tfrac{1}{2})\ ,\ f(-\tfrac{1}{2})\right)'.
\end{aligned}$$

Then, going through each term of (A.12) we have:

$$
\begin{aligned}
|A| \cdot (1\ ,\ 1)(\nabla A^\star) b\,\big|_{\varepsilon=0} &= -\big(f(-\tfrac{1}{2})\ ,\ f(\tfrac{1}{2})\big)\ b\,\big|_{\varepsilon=0} \\
&= -f(-\tfrac{1}{2})F(\tfrac{1}{2}) - f(\tfrac{1}{2})(1 - F(-\tfrac{1}{2})), \\
|A| \cdot (1\ ,\ 1)A^\star(\nabla b) &= -\tfrac{1}{2}(1-2p)\left(f(\tfrac{1}{2}) + f(-\tfrac{1}{2})\right), \\
-(\nabla |A|)\,(1\ ,\ 1)A^\star b &= (1-p)\big(f(\tfrac{1}{2}) + f(-\tfrac{1}{2})\big)\big(F(\tfrac{1}{2}) + 1 - F(-\tfrac{1}{2})\big).
\end{aligned}
$$

Gathering all terms, we obtain the following expression for the derivative:

$$
\begin{aligned}
&\tfrac{1}{4}\big(f(\tfrac{1}{2}) - f(-\tfrac{1}{2})\big) \quad + \quad \tfrac{1}{2}p\left(f(-\tfrac{1}{2})F(\tfrac{1}{2}) - f(\tfrac{1}{2})F(-\tfrac{1}{2})\right) \\
&\qquad + \quad \tfrac{1}{2}(p-1)\left(f(\tfrac{1}{2})F(\tfrac{1}{2}) - f(-\tfrac{1}{2})F(-\tfrac{1}{2})\right).
\end{aligned}
$$

Under symmetry of f around zero, we have $f(-\frac{1}{2}) = f(\frac{1}{2})$ and $F(-\frac{1}{2}) = 1 - F(\frac{1}{2})$. The expression then simplifies to

$$
\begin{aligned}
&\tfrac{1}{2}p\ f(\tfrac{1}{2})\left(F(\tfrac{1}{2}) - F(-\tfrac{1}{2})\right) + \tfrac{1}{2}(p-1)\ f(\tfrac{1}{2})\left(F(\tfrac{1}{2}) - F(-\tfrac{1}{2})\right) \quad = \\
&\tfrac{1}{2}p\ f(\tfrac{1}{2})\left(2F(\tfrac{1}{2}) - 1\right) + \tfrac{1}{2}(p-1)\ f(\tfrac{1}{2})\left(2F(\tfrac{1}{2}) - 1\right) \quad = \\
&-\tfrac{1}{2}(1-2p)\ f(\tfrac{1}{2})\left(2F(\tfrac{1}{2}) - 1\right).
\end{aligned}
$$

□

B Additional figures and tables

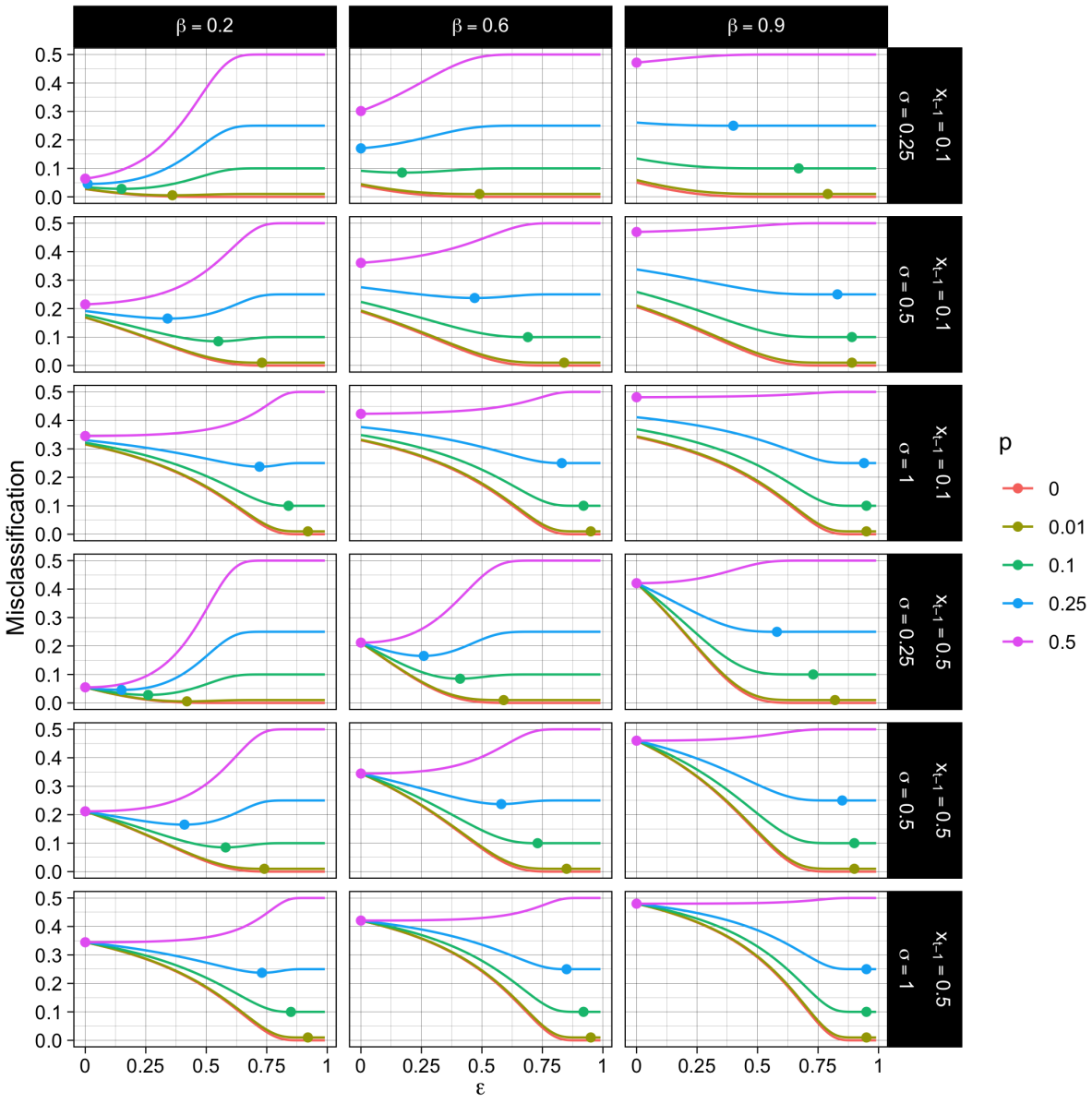


Figure B.1: 1-step-ahead misclassification rate using (11) and $c_{t-1} = 0$. The minimum of each curve is marked by a dot. The misclassification rate still presents a minimum, as in the case where $\beta = 0$, and the minimum misclassification probability is usually realized at non-trivial values of ε.

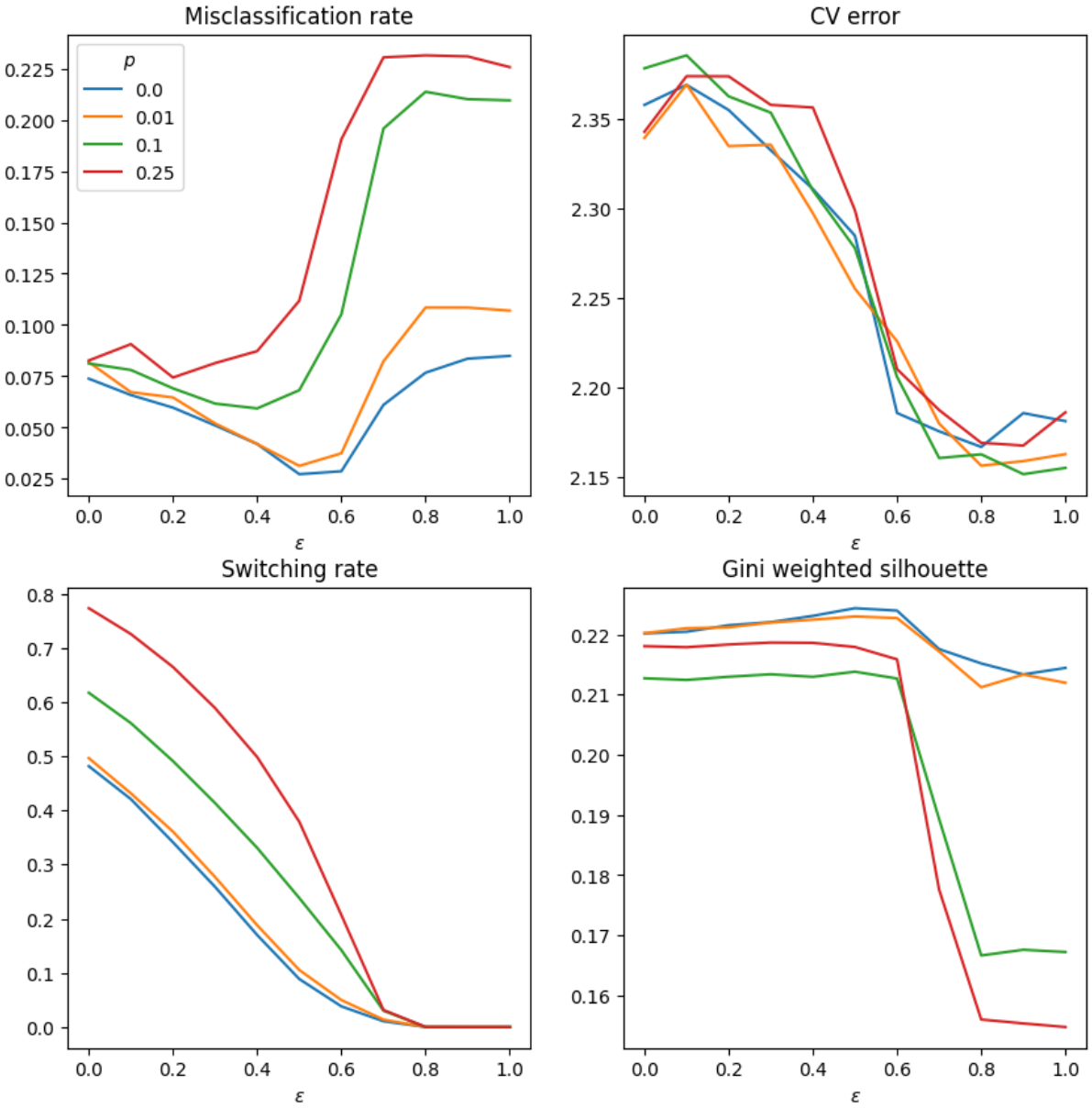


Figure B.2: Simulation results for four values of p. Half-variance setting.

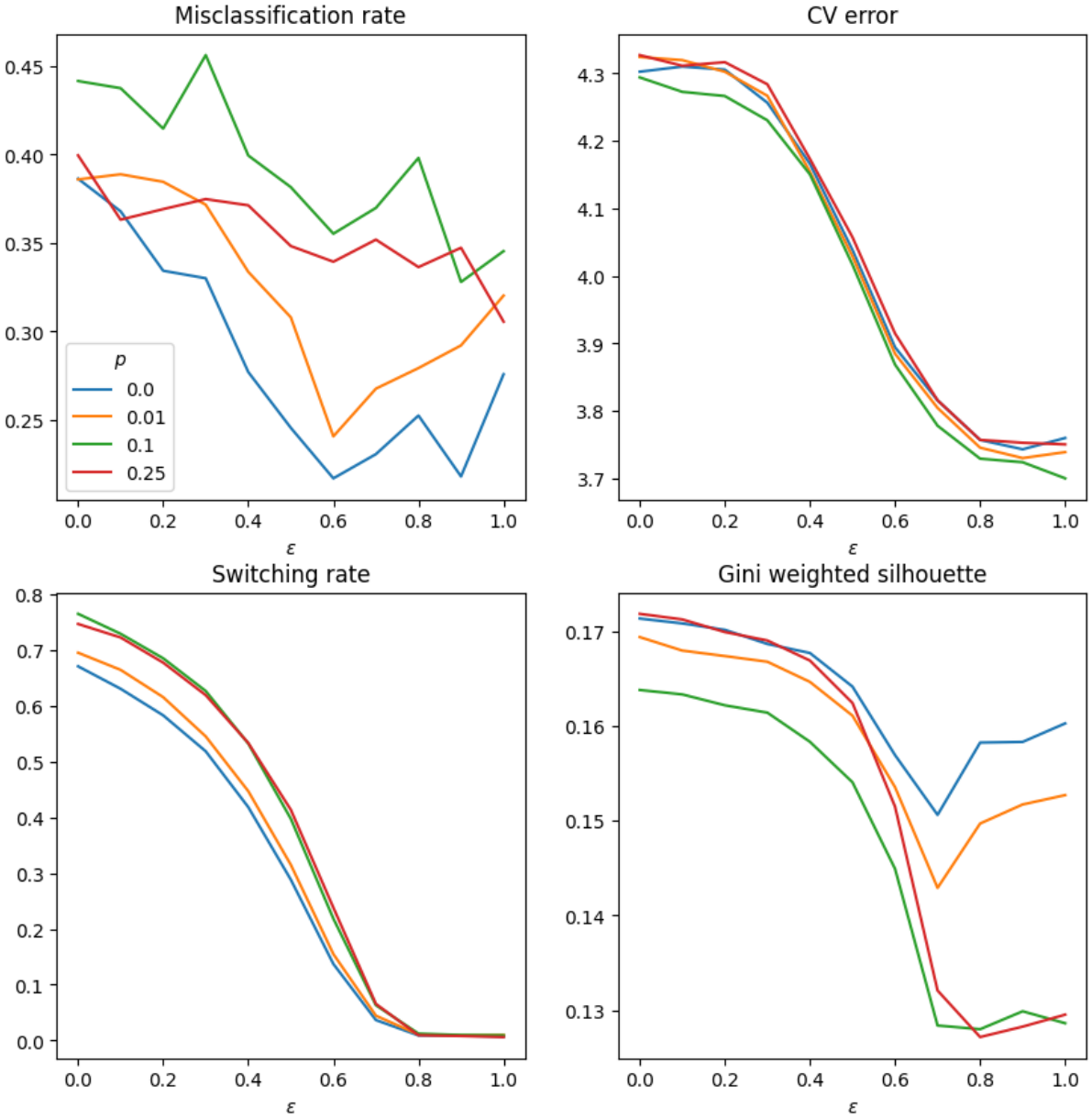


Figure B.3: Simulation results for four values of p. Benchmark setting. The number of clusters K_t may vary between 2 and 4, while the true number is always 2.

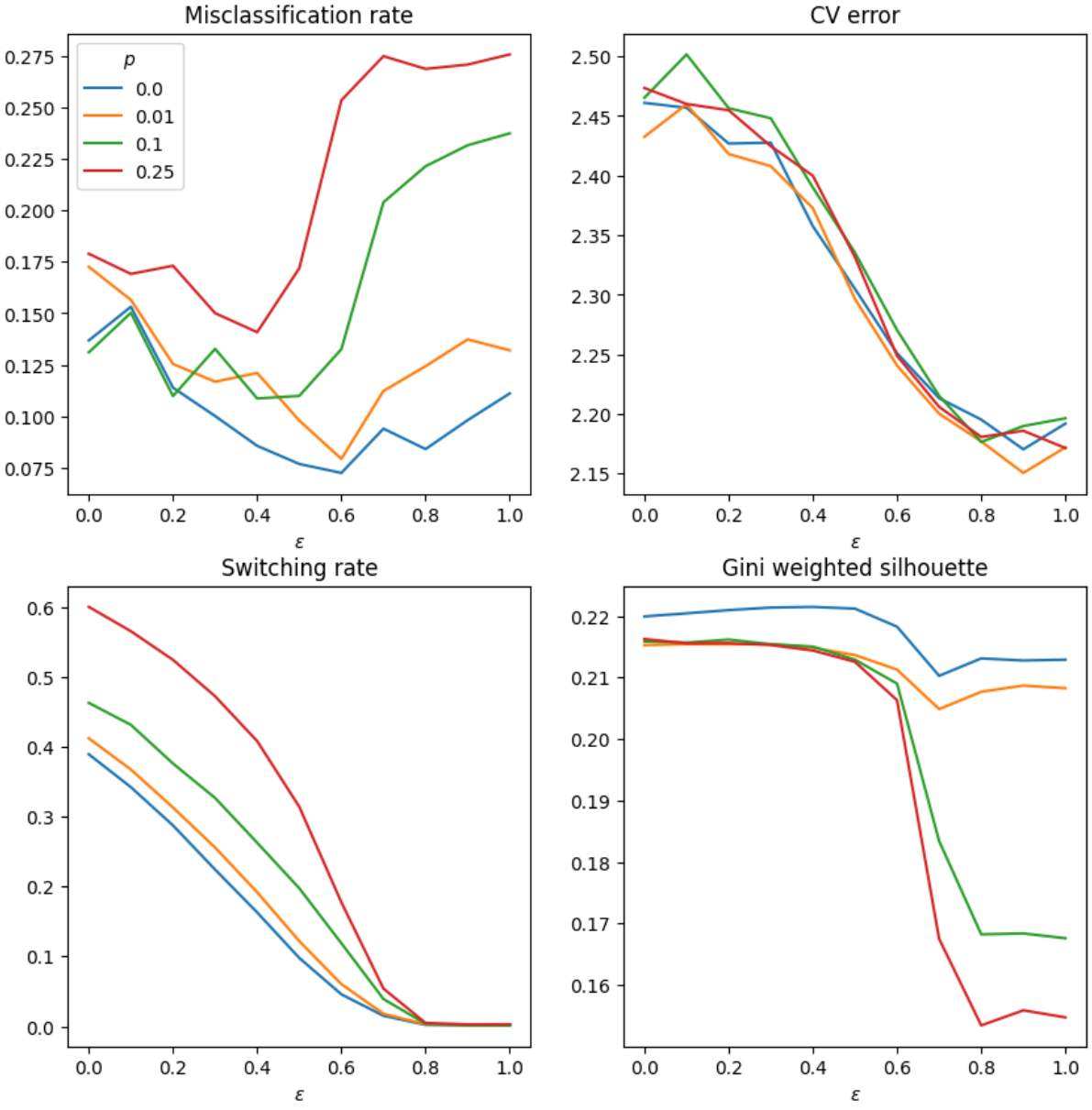


Figure B.4: Simulation results for four values of p. Half-variance setting. The number of clusters K_t may vary between 2 and 4, while the true number is always 2.

Table B.1: Cluster assignments for $K \equiv 4$ and $\varepsilon = 0.45$.

Name	2010	2012	2016	2020
Achmea Schadeverzekeringen NV	3	3	3	3
Allianz SE	3	3	3	3
Alte Leipziger	4	4	4	4
Assicurazioni Generali Spa	3	3	3	3
Covea	3	3	3	3
Credit Agricole Assurances	3	3	3	3
Danica Pension Livsforsikringsaktieselskab	2	2	2	2
Fidelidade - Companhia De Seguros SA	3	3	2	2
Gjensidige Forsikring Asa	3	3	3	3
Groupe Des Assurances Credit Mutuel SA	2	2	2	2
Hannover Re AG	1	1	1	1
KBC Verzekeringen	2	2	2	2
Livforsakringsbolaget Skandia, Omsesidigt	2	2	2	2
Mapfre SA	3	3	3	3
Munich Re AG	1	1	1	1
Nn Group NV	2	2	2	4
Pfa Holding AS	2	2	2	2
Pohjola Vakuutus OY	3	3	3	3
Powszechny Zaklad Ubezpieczen SA	3	3	3	3
R+V Versicherung AG	4	4	4	4
Sampo Oyj	3	3	3	3
Swiss Re AG	1	1	1	1
Ethniki Hellenic General Insurance Co. SA	3	3	3	3
Unipol Gruppo Spa	3	3	3	3
Vidacaixa Sa De Seguros Y Reaseguros	2	2	2	2
Vienna Insurance Group AG	3	3	3	3
Zavarovalnica Triglav	2	2	2	2
Zurich Insurance Group AG	3	3	3	3