

12.5 Dissimilarity-based multivariate control chart

Essential steps

Suppose we have an $(N \times p)$ data matrix, $\mathbf{Y}$, and we can capture the important relationships among the N sampling units in this matrix by calculating some chosen dissimilarity measure (e.g., Bray-Curtis) to yield an $(N \times N)$ dissimilarity matrix, $\mathbf{D}$. How can we create a control-chart from this? We may consider doing the following:

1. From matrix $\mathbf{D}$, obtain a set of *ordination axes*, held in an $(N \times m)$ matrix $\mathbf{Q}$, which adequately represent the inter-point relationships given in $\mathbf{D}$, but in a Euclidean space of dimension m .
2. Assume the 'in-control' samples arise from a multivariate distribution $\mathcal{D}$. Calculate a modification of Hotelling's T^2 criterion directly, using $\mathbf{Q}$ instead of $\mathbf{Y}$ (see the description of the modified test-criterion below).
3. Determine the upper control-chart limit (U_{CL}) in one of two ways:
 - Parametrically (assuming $\mathcal{D}$ is approximately multivariate normal); or
 - Non-parametrically (using a permutation procedure, hence distribution-free).

Taking the above steps will yield a dissimilarity-based control chart, yet which retains a useful desired property of classical multivariate control charts; namely, not only the *distance* from the 'in-control' centroid, but also the *direction* of a new point's position relative to that centroid will matter.

Description of test criterion

Consider the comparison of a multivariate sample obtained at a given time-point, t , by reference to a set of n_c in-control samples. Let $\mathbf{D} = \{d_{ii'}\}$ consist of the dissimilarities between every pair (i, i') of multivariate samples $(i = 1, \dots, n_c, t)$ and $(i' = 1, \dots, n_c, t)$, and so $\mathbf{D}$ is a matrix of dimension $((n_c + 1) \times (n_c + 1))$.

From $\mathbf{D}$, we do an ordination on the full set of $(n_c + 1)$ sampled time-points to generate $\mathbf{Q}$, a set of m ordination axes. Let the sample under test for time-point t , be an m -length vector in matrix $\mathbf{Q}$ denoted by $\mathbf{q}_t$. Furthermore, let $\mathbf{Q}_c$ denote the ordination positions for only the remaining n_c in-control (reference) samples (omitting $\mathbf{q}_t$). Also, let the m -vector of mean values calculated using all of the reference samples be denoted by $\bar{\mathbf{q}}_c$. We shall assume the in-control samples $\mathbf{q}_i$ for $(i = 1, \dots, n_c)$ arise from a common multivariate distribution $\mathcal{D}$, with mean $E(\mathbf{q}_i) = \boldsymbol{\mu}_c$ and variance-covariance $\text{Var}(\mathbf{q}_i) = \boldsymbol{\Sigma}_c$.

To obtain our test criterion, we begin by calculating:

$$z_t = \sqrt{\frac{n_c}{(n_c+1)}} (\mathbf{q}_t - \bar{\mathbf{q}}_c)$$

This standardisation, including the multiplier, ensures that, if the null hypothesis is true and $\mathbf{q}_t$ also arises from $\mathcal{D}$, then $E(z_t) = 0$ and $\text{Var}(z_t) = \Sigma_c$.

We then define our new control-chart test-criterion as:

$$T^2_t = \frac{1}{m} \mathbf{z}_t^T \mathbf{S}_{Q_c}^{-1} \mathbf{z}_t$$

where $\mathbf{S}_{Q_c}^{-1}$ is the classical unbiased estimator of the variance-covariance matrix calculated using only the in-control samples of matrix $\mathbf{Q}_c$. If desired, [shrinkage](#) can also be applied here, in which case $\mathbf{S}_{Q_c}^{-1}$ will be replaced by $\mathbf{W}_{Q_c}^{-1}$

There are several ways that this charting criterion differs from Hotelling's criterion used in classical multivariate control charts. First, the ordination will be done afresh for each successive time-point under test. Thus, both the observed value of T^2_t and also its distribution will rather naturally depend on t . Furthermore, we expect that the value of m (i.e., the number of dimensions required by the ordination method to accommodate an increasing number of sampling points) will also increase over time. Hence, to make the control chart easier to read, our criterion includes, for plotting purposes, the multiplier $\{1 \text{ over } m\}$, so that values are expressed as a standardised T^2 distance per number of dimensions. However, importantly, the value of m , once chosen, does not change *within* a given time-point, so inclusion of the multiplier will not affect comparisons of the observed value of T^2_t with its null distribution in any material way.

Ordination methods and choice of m

There are several potentially suitable ordination methods that may be used to produce $\mathbf{Q}$, including principal coordinate analysis (PCO; [Gower \(1966\)](#)), or a multi-dimensional scaling method that is metric (mMDS; [Sammon \(1969\)](#), [Borg & Groenen \(2005\)](#)), threshold metric (tmMDS; [Clarke et al. \(2014\)](#)) or non-metric (nMDS; [Kruskal & Wish \(1978\)](#)).

What we are after is a set of m coordinate axes, represented here by an $((n_c+1) \times m)$ matrix $\mathbf{Q}$, whose Euclidean inter-point distances $\{e_{ii'}\}$ match the original dissimilarities $\{d_{ii'}\}$ (or their ranks) extremely well. A natural question is: how closely should ordination distances match original distances? In other words: how many ordination axes shall we use to represent dissimilarities in Euclidean space (i.e., what value shall we choose for m)? It is important to retain as much original information, natural variation and complexity inherent in the original system as possible, but without including redundancies or distortions.

If **PCO** is used, it would be useful to exclude: (i) any PCO axes corresponding to positive eigenvalues that occur as an artefact to inflate the total variance of the system; and (ii) any PCO axes corresponding to negative eigenvalues, if any ([McArdle & Anderson \(2001\)](#)). Thus, we could

choose to use a maximum value of m that will still maintain the relationship:

$$\sum_{i \neq i'} e_{ii'}^2 \leq \sum_{i \neq i'} d_{ii'}^2$$

That is, we could choose to maximise m such that

$$100 \times \sum_{i \neq i'} e_{ii'}^2 / \sum_{i \neq i'} d_{ii'}^2 \leq b$$

where $b = 100$ percent. However, we may alternatively choose $b = 90$ percent or 80 percent, etc., in an effort to reduce noise. A threshold value of $b = 80$ percent would seem reasonable, but the choice here rests with the end-user.

For **metric MDS** (mMDS), one might choose m so that the Pearson matrix correlation ($r_{e,d}$) between the Euclidean distances in the m -dimensional MDS space $\{e_{ii'}\}$ and the original dissimilarities $\{d_{ii'}\}$ exceeds some threshold value (e.g., $r_{e,d} \geq 0.99$). The latter criterion was suggested by [Clarke et al. \(2014\)](#) in the context of performing bootstrap averaging in an m -dimensional metric MDS space (see [chapter 18](#) therein). A similar rationale and agenda is desirable here – we wish to have a set of Euclidean axes that avoids inappropriate noise and redundancies (it is sensible to avoid the nonsensical conclusion that every single replicate might be considered an outlier), but nevertheless retains core information captured by the inter-point dissimilarities. A threshold value for this matrix correlation of $r_{e,d} \geq 0.95$ would also seem reasonable, but the choice here is, once again, left to the end-user.

It is also possible to use **non-metric MDS** (nMDS) here. In this case, the matrix correlation is constructed using the Spearman rank correlation coefficient ($\rho_{e,d}$), rather than the Pearson correlation coefficient, but all else is the same. In practice, however, the use of either threshold metric or metric MDS would seem a better option here than to use non-metric MDS. First of all, the latter retains only rank-order relationships of dissimilarities. However, a control chart, by its very nature, is designed to quantify the distance from a new point to a distribution of prior points. In this context, the preservation of rank dissimilarities only (via nMDS) would not be expected to provide consistent results. Furthermore, nMDS has the potential to yield [degenerate solutions](#) (of low stress) specifically when there is one (or more) outliers (or genuine splits in the data), which further suggests it would not be the best choice to use here.

Threshold metric MDS (tmMDS) differs only from metric MDS in permitting a non-zero intercept in the construction of the Shepard diagram, which in practice means that two samples that occupy the same position in the tmMDS may be interpreted as yet to differ by some threshold amount (i.e., the value of the non-zero intercept). This does not pose any obvious problem in the context of constructing a control chart, and as tmMDS also tends to achieve lower stress for an equal choice of m by comparison with metric MDS, we consider it to be a good (default) choice for creating ordination axes that can be used routinely to build control charts.

Parametric upper control-chart limit

It may be very reasonable to assume that the distribution of samples $\mathscr{D}$ under a true null hypothesis H_0 in the ordination space $\mathbf{Q}$ is approximately multivariate normal. Even if

the original variables in $\mathbf{Y}$ are not the least bit normally distributed (e.g., they may be zero-inflated, overdispersed, aggregated and/or have strong mean-variance relationships, etc.), the distribution of samples in the space of the resemblance measure, whose inter-point patterns are captured by $\mathbf{D}$, will likely be quite even, with few outliers, if H_0 is indeed true.

Thus, noting that our modified criterion T^2_t (above) differs from the classical Hotelling T^2 statistic only by a factor of $\frac{1}{m} \times \frac{n_c}{(n_c+1)}$, its distribution under a true null hypothesis is:

$$T^2_t \sim \frac{(n_c - 1)}{(n_c - m)} F_{m, (n_c - m)}$$

Accordingly, the upper control-chart limit at a chosen significance level, α , is therefore given by

$U_{CL} = \frac{(n_c - 1)}{(n_c - m)} Q_{(1-\alpha)}[F_{m, (n_c - m)}]$ This limit will likely be different for different values of t , because the value of m for the ordination (created anew for each value of t) may differ. If a constant value for m is chosen for the entire control chart, and the value of n_c also does not change with t (this is true for the baseline and moving-window types of control charts), then the value of the parametric U_{CL} will remain constant as well.

Non-parametric upper control-chart limit

We may, alternatively, consider a non-parametric approach. Here, we shall assert only that the multivariate data points follow a stochastic process over time. We propose using a permutation procedure to obtain the upper control chart limit at any particular time-point t .

Under a true null hypothesis, all $\mathbf{q}_i$, $i = 1, \dots, n_c$, arise from a common distribution $\mathscr{D}$. We add to this the notion of exchangeability through time; specifically, all of the 'in-control' points in the reference set, i.e., the $\mathbf{q}_i$, could appear in any order relative to one another, up to and including time n_c . Under this assumption, we can permute the (sample) rows of $\mathbf{Q}_c$ to obtain $\mathbf{Q}^*_c$ and consider the last observation (in the n_c^{th} row of $\mathbf{Q}^*_c$) to be a 'new point' for the test, but where we know that H_0 is actually true. We calculate $T^{2*}_{n_c}$, which is the value of the proposed control-chart test-statistic that compares this last observation to the distribution of the other (n_c-1) in-control points.

We repeat this permutation procedure many times to get an empirical distribution of values for $T^{2*}_{n_c}$. Note that the number of unique values we can get under permutation here is actually severely limited by n_c . Each sample in the original reference set can only take on the role of being the 'tested' point once, so there are only n_c unique values of $T^{2*}_{n_c}$ possible under permutation. Nevertheless, we can calculate a permutation-based upper control-chart limit as the $(1-\alpha)$ percentile on that empirical distribution, specifically: $U_{CL} = Q_{(1-\alpha)}[T^{2*}_{n_c}]$ Note that this non-parametric upper control-chart limit will be different for every value of t .