**Supplementary Information :**

**Figure S1 – Colour Version of Weak Aggregate (Main Text – Figure 2)**



**Figure S2 – Colour Version of Strong Aggregate (Main Text – Figure 3)**



**Application of Singular Valued Decomposition (SVD) to Derive Strong Aggregate**

Suppose that, as in the main text, the distinction between disease states in the neighbourhood of prototypes C and D are clinically relevant. Using the positive, negative and general symptom domain scores for each patient, a 3-dimensional space is formed, shown in Figure 2 and 3 of the main text, and above, in colour form.

We first define the *M* x *N* matrix, **Y**, where rows, *M*, are variables measuring clinical state e.g. PANSS domains (positive, negative and general respectively) and the columns are values of each variable for the 1459 patients. We then append to **Y** the prototypes A, B, C and D resulting in *N* = 1463.

An SVD (Strang, 2004) decomposes **Y** into three matrices:

$$Y=UΣV^{T}$$

Where $U$ and $V$ are square matrices (the so-called left and right singular vectors) whose columns are the orthogonal eigenvectors of the matrix $YY^{T}$ and $Y^{T}Y$ respectively. Elements (*i*,*j*) of $Y^{T}Y$represent a measure of similarity between patients *i* and *j* based on co-occurrence of their PANSS domains scores (positive, negative and general). $Σ$ is the diagonal matrix of singular values of $Y$.

We can now examine the pattern of similarity between patients and prototypes by using only the columns of $V$ representing the 3 largest singular values - resulting in a matrix of size 1463 x 3, which can be visualised as shown in Figure S3.

**Figure S3**

By inspection, we find that the prototypes of clinical relevance that we wish to emphasise (see main text), C and D, are most separated along the axis V3 (corresponding to the 3rd singular value). We label this the *best* or *optimal* dimension *V\**. To form a univariate score, we simply retrieve the position of each of the 1459 points (for each patient) on *V\**. To define our strong aggregate (in the traditional sense of a univariate outcome), we can then use either a hard or soft threshold. In the former, we chose a fixed point along *V\**, and all patients higher/lower are assigned a binary outcome (which could represent for example, response or non-response). Alternatively, we used a ‘soft-threshold’, assigning each patients’ strong aggregate value according to a smooth function over their values along *V\** (Figure 3 in the main text, shown in colour above as Figure S2).

**Using SVD For Matching / Retrieving Patients**

In the Discussions and Conclusions of the main text, we suggested that beyond defining strong aggregates, the formulation of SVD has utility for ‘matching’ patients. In this application - as it is in information retrieval and latent semantic indexing (Manning, Raghavan and Shutze, 2008) – given an SVD of some set of patients represented by $Y$defined as before (e.g. rows correspond to variables measuring clinical state via an instrument such as PANSS items or domains, with columns again representing individual patients), a new patient becomes a ‘query’ for which we seek similar patients and measure the proximity of the new patient to prototypes.

Assume an SVD has been constructed in exactly the same way as before, but now, we retain some number *k* of the singular values (i.e. rather than 3, we might opt for larger *k*). This results in the reduced-rank representation of $Y$as follows:

$$Y\_{k}=U\_{k}Σ\_{k}V\_{k}^{T}$$

$Y\_{k}$ is now a lower-dimensional approximation of the original $Y$that, by virtue of retaining only a subset of the largest *k* singular-values, reduces redundancy and noise, emphasising the dominant similarities between patients’ clinical states. Any new patient $y^{'}$ represented by their state in the original space $Y$can be queried for similarity to other patients and prototypes by first computing their projection into the lower dimensional space $Y\_{k}$ :

$$y\_{k}^{'}= Σ\_{k}^{-1}U\_{k}^{T}y^{'}$$

Then, comparison to other patients (and prototypes) proceeds by measuring the cosine similarity between $y\_{k}^{'}$ and columns of $V\_{k}^{T}$ – see (Manning, Raghavan and Shutze, 2008) for further details and demonstration.

Manning, C. D., Raghavan, P. and Shutze, H. (2008) *An Introduction to Information Retrieval*. Cambridge University Press. doi: 10.1109/LPT.2009.2020494.

Strang, G. (2004) *Linear Algebra and Its Applications*. 4th edn. WB Saunders. doi: 10.2307/2003783.