ggvis help

Andreas Buja and Deborah F. Swayne

May 6, 2004

1 Introduction

ggvis uses multidimensional scaling (MDS) to map a set of objects to points in k-space such that
dissimilarity(object i, object distance(point i, point j)
This is made precise by minimizing the chosen stress function.
The parameters of the function are:

- D = target distance matrix This is either supplied as a matrix of dissimilarities or distances, or it uses the shortest-path metric if a discrete graph is given in terms of a links file.
- p = exponent of the power transformation of D
- l = Minkowski norm used to calculate the current interpoint distances ($l = 2$ corresponds to the Euclidean metric on k-space)
- r = residual power applied to the difference between the target distance D and the interpoint distances
- T = threshold above which dissimilarities are removed from the stress function

The value of the stress function is plotted as a number generally between zero and one.
Below the plot of the stress function is a barplot of the distribution of D^p. As you change p, this plot is updated. Use the left mouse button to set the threshold T used in the stress function: the solid bars are below the threshold and represent the distances used in MDS; the hollow bars correspond to pairs of points that are ignored.
Other parameters available on the control panel are

- Dimension (k): the dimensionality of the space into which the data are being mapped. The default is 3D. Examine your point configurations with 3D rotations or grand tours in ggobi.
- Stepsize: the size of the step taken to bring x_i and x_j closer together or to push them farther apart. You might start with a large stepsize and then reduce it once the structure begins to stabilize. You can also increase it suddenly in a form of interactive "simulated annealing," to jump out of a local minimum when you think there might be a lower minimum elsewhere.
- MDS within subgroups: If you have subgroups in the data that you would like to handle separately from one another, then brush them using ggobi. Then select this button and restart "Run MDS." Now only distances between points with the same glyph and color will be considered.

- MDS casewise: If you would like to investigate the behavior of a single case, or a small group of cases, then use ggobi to add persistent labels to those cases. Then select this button and restart "Run MDS." Now only distances between pairs of points including one of the labelled points will be considered. (This feature can be used in conjunction with brush groups. Then only distances between pairs of points with the same color and glyph AND within the same brush group will be considered.)

- Shepard diagram: Launch ggobi display containing a plot of D^p vs the configuration distances.

2 Optimization of the configuration

Since you were able to start up ggvis, begin by clicking "Run MDS" and see what happens: You should see the points in the ggobi window move and a curve evolve in a subwindow titled "Stress function" in the ggvis panel. The motion is generated by a gradient descent algorithm; the curve shows the criterion, called Stress.

Click again and the motion stops. Click one more time and leave "Run MDS" on in what follows. This allows you to immediately see the effects of various manipulations.

By default, ggvis performs gradient descent of a stress function for metric Kruskal-Shepard distance scaling.

3 Controls in the ggvis panel

3.1 First tabbed widget: Specify datasets, define D, and Run

- k = Dimension of the points, default: 3. For 3D viewing, choose in the ggobi console ViewMode → Rotation For 2D viewing, choose in the ggobi window ViewMode → XYPlot The dimension k can be as high as 12. For viewing 4D and higher, choose in ggobi ViewMode → 2D Tour

- Stepsize: The stress optimization uses a fixed-size gradient step. This slider permits controlling the step size in terms of a fraction of the size of the point configuration. Default: 0.02 (gradient sz = 2% config sz).

- Run MDS: Turn optimization on and off

- Step: an alternative to "Run MDS" if you wish to follow the gradient steps one by one.

3.2 Plot of the stress function

Stress function plot: shows the evolution of the stress values as optimization progresses. No interactive controls.

3.3 Key MDS parameters

- "Metric" versus "Nonmetric": choice of metric and nonmetric MDS

- "Krsk/Sh" versus "Classic": choice of Kruskal-Shepard distance scaling versus Torgersen-Gower dot-product scaling (also called classical MDS).

IMPORTANT for those who know about MDS: ggvis has a version of nonmetric classical (!) scaling. That is, the two pairs of choices above are in fact fully crossed. Nonmetric classical scaling is not very stable, but it works with good starting configurations.

[Technical remark: The idea is to perform dot-product scaling of $-D^2$ under the constraint that the configuration remains centered. This is equivalent to doubly-centering $-D^2$ and performing unconstrained dot-product scaling. The advantage of the former is that it can be made nonmetric by replacing $-D^2$ with an isotonic transform $f(-D)$ which can be estimated by regressing the configuration distances on$-D$.]

3.4 Data (D): Histogram, power, weights

- Histogram of transformed dissimilarities (bottom of the ggvis panel): Note the grips on either end of the horizontal axis! You can move them to trim large and small dissimilarities. The right grip trims large ones; the left grip small ones. This allows you to check their importance.
- p = Exponent of the power transformation of D in metric MDS. Playing with the slider allows you to interactively optimize stress. Also, see what happens as $p \to 0$ (scaling of a simplex).
- r = Weight parameter: The dissimilarities D can be weighted with $w = D^r$ in the stress function. Default: $r = 0$, equal weights.

3.5 Menubar: View, Reset

- View

  "Shepard Plot": starts up a separate ggobi window with the following variables:

  - d_{ij} = interpoint distances
  - $f(D_{ij})$ = transformed dissimilarities; metric: power; nonmetric: isotonic.
  - D_{ij} = raw dissimilarities
  - Res_{ij} = residuals $f(D_{ij})$ - d_{ij}
  - Wgt_{ij} = weights, informative only if r or wb is non-default i = the first index useful for checking the patterns j = the second index / of missing or omitted dissimilarities

  The number of dissimilarities can be large! The field below "Shepard Plot" tells you how many there are. For N¿500 you might want to be careful. Use "Select'n prob" with "Run MDS" off to get a random sample of manageable size.

- Reset

  - Reinit layout: Re-initialize the configuration to the initial positions.
  - Scramble layout
  - Reinit parameters

3.6 Final tabbed widget: distance, groups, sensitivity, constraints

beginitemize

- Distance

- p = exponent of the power transformation of D
- m = Minkowski norm: Used to calculate the current interpoint distances. Default: $m = 2$, the Euclidean metric on k-space.

- Groups
  - Click "Groups", a window for group selection will pop up. Click in the "Hide" column to remove and restore groups. Watch the effects on the point configuration.
  - Click on the menu underneath "MDS with Groups".
    * Select "Dists within groups": same as wb=(2,0) above, but implemented more efficiently without weights. Similar for "Dists between groups" (wb=(0,2)). Default: "Ignore groups" (wb=(1,1)).
    * Select "Dists from&in anchor": anchored MDS maps points within an anchor set by MDS (as usual), but it maps the points outside the anchor set with the dissimilarities w.r.t. the anchor set (hence the term).
      The anchor set consists of the points with the glyph that appears to the right of "MDS with Groups". You can change the anchor set by selecting "View:..." -¿ "Move Points" in the ggobi window. Click MIDDLE on a point, and the points with that color glyph become the new anchor set.
    * Select "Dists from anchor": same, except the anchor points remain fixed. This allows you to mess with the anchor points manually by dragging them around with the LEFT mouse button depressed.
  - wb = Within/between parameter: If there are color/glyph groups, the parameter allows one to differentially weight dissimilarities that link points within groups versus between groups. Extremes: wb=(2,0): Use only within groups, remove between groups; wb=(0,2): Remove within groups, use only between groups. Default: wb=(1,1): Weight within and between groups equally.
    Color/glyph groups can be provided in input or created interactively with ggobi's Brush mode.

- Sensitivity
  - Selection probability: random subselection of dissimilarities; a form of stability check. Implementation: a uniform [0,1] random number is generated for each dissimilarity, which is included in the stress function if the random number is below the selection probability read from the slider. The "New" button allows you to create a new set of random numbers. Select for example a selection probability 0.9 and click "New" repeatedly. This will show how stable the configuration is under removal of roughly 1 in 10 dissimilarities. Default: 1.0 (all dissimilarities included).
  - Perturbation: perturb the configuration with normal random vectors. The slider value s determines the relative fractions of configuration and perturbation: (1-s)*Config + s*Random. Extremes: s=1.0: pure random, i.e., a random start (default). s=0.0: pure configuration, i.e., no perturbation. Click "New" to initiate a perturbation.

- Constraints

4 Formulas

Kruskal-Shepard Distance Scaling

$$\text{STRESS}_D(\mathbf{x}_1,...,\mathbf{x}_N) \;=\; \Big(1-\cos^2\Big)^{1/2}$$

$$\cos^2 \;=\; \frac{\Big(\sum_{(i,j)\in I}\; w_{i,j}\cdot f(D_{i,j})\cdot \|\mathbf{x}_i-\mathbf{x}_j\|_m^q\Big)^2}{\Big(\sum_{(i,j)\in I}\; w_{i,j}\cdot f(D_{i,j})^2\Big)\Big(\sum_{(i,j)\in I}\; w_{i,j}\cdot \|\mathbf{x}_i-\mathbf{x}_j\|_m^2\Big)}$$

$D_{i,j}\in \mathbb{R},\;\geq 0,\; N\times N$ matrix of dissimilarity data

$$f(D_{i,j}) \;=\; \begin{cases} D_{i,j}^p\;, & \text{for metric MDS} \\ \text{Isotonic}(D_{i,j})\;, & \text{for nonmetric MDS}\end{cases}$$

$0\leq p\leq 6,\;$ default : 1 (no transformation)
Isotonic = monotone transformation estimated with isotonic regression

$\mathbf{x}_1,...,\mathbf{x}_N\in \mathbb{R}^k,\;$ configuration points
$1\leq k\leq 12,\;$ default : 3 (use 3D Rotations or Grand Tour)

$$\|\mathbf{x}_i-\mathbf{x}_j\|_m^q \;=\; \Big(\sum_{\nu=1,...,k}|x_{i,\nu}-x_{j,\nu}|^m\Big)^{q/m},\qquad \text{configuration distances, } (..)^q$$

$1\leq m\leq 6,\;$ default : 2 (Euclidean; $m=1:$ Cityblock)
$0\leq q\leq 6,\;$ default : 1 (STRESS; $q=2:$ SSTRESS)

$$w_{i,j} \;=\; D_{i,j}^r\cdot\begin{cases} w\;, & \text{if color/glyph of } i,j \text{ is same} \\ (2-w)\;, & \text{if color/glyph of } i,j \text{ is different}\end{cases}$$

$-4\leq r\leq +4,\;$ default : 0 (equal weights; $r=-1:$ Sammons)
$0\leq w\leq 2,\;$ default : 1 (ignore grps; $w=2:$ within; $w=0:$ between grps)

$$I \;=\; \{\;(i,j)\mid i\neq j,\; D_{i,j}\neq NA,\; T_0<D_{i,j}<T_1,\;\; \text{Runif}(i,j)<\alpha,\;\ldots\;\}$$

$0\leq T_0\leq T_1,\;$ thresholds, defaults : $0,\;\infty$
Runif = uniform random numbers $\in[0,1]$
α = selection probability, default : 1
... = conditions based on color/glyph groups

Torgerson-Gower Dot-Product Scaling

$$\text{STRAIN}_D(\mathbf{x}_1,...,\mathbf{x}_N) \;=\; \Big(1-\cos^2\Big)^{1/2}$$

$$\cos^2 \;=\; \frac{\Big(\sum_{(i,j)\in I}\; w_{i,j}\cdot f(-D_{i,j}^2)\cdot \langle\mathbf{x}_i,\mathbf{x}_j\rangle\Big)^2}{\Big(\sum_{(i,j)\in I}\; w_{i,j}\cdot f(-D_{i,j}^2)\Big)\Big(\sum_{(i,j)\in I}\; w_{i,j}\cdot \langle\mathbf{x}_i,\mathbf{x}_j\rangle\Big)}$$

$D_{i,j} \in \mathbb{R},\ \geq 0,\ N \times N$ matrix of dissimilarity data

$$f(-D_{i,j}^2) \ = \ \begin{cases} -D_{i,j}^{2p}\ , & \text{for metric MDS} \\ \text{Isotonic}(-D_{i,j})\ , & \text{for nonmetric MDS} \end{cases}$$

$0 \leq p \leq 6,$ default : 1 (no transformation)
Isotonic = monotone transformation estimated
with isotonic regression

$\mathbf{x}_1, ..., \mathbf{x}_N \in \mathbb{R}^k,$ configuration points, constrained to $\sum \mathbf{x}_i = 0$
$1 \leq k \leq 12,$ default : 3 (use 3D Rotations or Grand Tour)

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle \ = \ \sum_{\nu=1,...,k} x_{i,\nu} \cdot x_{j,\nu}, \qquad \text{configuration dot products,}$$

$$w_{i,j} \ = \ D_{i,j}^r \cdot \begin{cases} w\ , & \text{if color/glyph of } i,j \text{ is same} \\ (2-w)\ , & \text{if color/glyph of } i,j \text{ is different} \end{cases}$$

$-4 \leq r \leq +4,$ default : 0 (equal weights)
$0 \leq w \leq 2,$ default : 1 (ignore grps; $w = 2:$ within; $w = 0:$ between grps)

$$I \ = \ \{\ (i,j) \mid i \neq j,\ D_{i,j} \neq NA,\ T_0 < D_{i,j} < T_1,\ \ \text{Runif}(i,j) < \alpha,\ \ ...\ \}$$

$0 \leq T_0 \leq T_1,$ thresholds, defaults : $0,\ \infty$
Runif = uniform random numbers $\in [0,1]$
α = selection probability, default : 1
... = conditions based on color/glyph groups