

Balancing constriction and expansion for domain-specific grammars

Jason Cory Brunson*

Laboratory for Systems Medicine, University of Florida

ABSTRACT

Graphical grammars distinguish from heuristic typologies in constituting closed formal systems whose product diversity arises from recombination rather than from *de novo* redesign. In practice, the scope of an implementation is constrained both by what options have been implemented for each component (e.g. the polar coordination, the grouped count statistic, and the rectangle geometric element) and by what combinations are enabled (these three yielding a pie graphic). In the interest of flexibility, implementations are generally permissive of syntactically valid combinations. However, within some domains, universal or conditional rules may be in place to ensure that graphics faithfully communicate important properties of the input data, such that permitting violations would compromise their integrity. I will present geometric data analysis as a motivating case study for this situation. Visualizations of point clouds, and especially the use of biplots, relies universally on a unit aspect ratio and conditionally on coordinate standardization for plot elements to approximate geometric relationships among data elements. Rather than a limitation of graphical grammar, these should be viewed as opportunities to explore the phrase spaces of different graphical domains. Indeed, the geometric constraints above enable the meaningful use of two families of statistical transformation—projections of constructions in higher dimensions and constructions in relation to referent points—that would otherwise produce uninterpretable graphics. This leads me to argue for a change in language from Wilkinson’s assertion that, while a grammatical system can produce hideous graphics, it cannot produce meaningless ones. I suggest that meaningfulness and meaninglessness are not mathematical properties secured by formal rules but semantic properties vested with cultural significance, and that graphical grammars must adapt to the culture of a domain in order to serve as trustworthy tools for communicating information within it.

Index Terms: Grammar of graphics, geometric data analysis, singular value decomposition, biplot.

1 GRAMMAR OF GRAPHICS

In his book introducing the grammar of graphics (GoG), Wilkinson [6] emphasizes the value not only to rigor but also to extensibility of taking a grammatical, in opposition to a typological, approach to statistical graphics:

The remarkable consequence of building a closed formal system is that, while it solves more complex applied problems, it can appear more adaptive to the user. Paradoxically, closed systems often behave more “openly” than open systems.

This flexibility assumes and relies on a permissiveness toward the “phrase space” of graphics that may be realized through syntactically valid combinations. Indeed, Wilkinson asserts that, while his system can produce many hideous graphics, it cannot produce

meaningless ones. On my reading, the primary focus, in the interest of meaningfulness, is on the mathematical determinism connecting the data to the geometric elements of a graphic—ultimately defined as a visual representation of the abstract *graph* of a mathematical function.

To be sure, Wilkinson details how meaningful graphics must also account for human perceptual tendencies and cultural norms. In popular implementations, led by Wickham [5], most directives toward good practice take the form of defaults (e.g. different color palettes for different variable types) and warnings (e.g. that opacity should not encode a discrete variable). Wilkinson also posits certain near-universal constraints, for example that the plotting window should be sized to fit the input data, even if only summary elements are visualized.

What is missing in this account, I believe, is a reckoning with dialect: Certain disciplines manage structured data that imposes specialized constraints on graphical representations, so that extending the grammatical approach to produce them requires constraining the combinatorial freedom it makes possible. Rather than choosing between uninhibited extensions that enable misleading graphics and a duplicative re-implementation for each domain, we should design and adopt principles of fidelity between important properties of the data to the graphical design patterns that encode them.

2 GEOMETRIC DATA ANALYSIS

GDA is a diverse family of techniques for analyzing data represented as a *point cloud*—a finite multisubset $X \subset (\mathbb{R}^d, L^2)$ —and for transforming other data types into point clouds for analysis. The points $x_i \in X$ may be attributed *masses* m_i , and the centroid of X is taken as the origin, so that the point *inertias* $m_i \|\vec{x}_i\|^2$ sum to zero. A foundational step in GDA is the decomposition of inertia via singular value decomposition (SVD): Taking $X_{n \times m}$ to be centered and $r \leq \min(n, m)$, we obtain

$$X = U_{n \times r} \Sigma_{r \times r} V_{m \times r}^\top$$

with coordinatewise inertias $\text{diag}(\Sigma) = (\sigma_1, \dots, \sigma_r)$. The point cloud $F = U\Sigma$ is a rotation of an optimal approximation to X , with equality when $r = \min(n, m)$; the column vectors $\vec{v}_1, \dots, \vec{v}_r$ of V each maximize $\text{Var}(\text{proj}_{\vec{v}_i} F_{i-1})$, where F_{i-1} is the **normal space/orthogonal complement** to the span $\langle \vec{v}_1, \dots, \vec{v}_{i-1} \rangle$ (starting with $F_0 = F$); and each $\text{Var}(\text{proj}_{\vec{v}_i} F_{i-1}) = \sigma_i^2$. SVD thus obtains an ordered set of *artificial coordinates* for the proxy point cloud F .

Two of the most widely used GDA techniques are principal components analysis (PCA), which obtains X by centering and scaling numeric case-variable data, and correspondence analysis (CA), which obtains a matrix of standardized residuals from bivariate count data. These techniques are called *ordination* as opposed to regression (being unsupervised rather than supervised) and *gradient analysis* as opposed to cluster analysis (yielding continuous rather than discrete artificial variables). The artificial coordinates (*principal coordinates* along *principal axes*) are not specified from a model but derived from data; because they carry no intrinsic meaning, they are sometimes called *scaffolding*.

The SVD supports a unique statistical graphic, the *biplot*, introduced by Gabriel [1]. Because the matrices U and V share their r artificial column variables, row markers for both (proxy row and column markers for X) can be overlaid on those axes. When F

*e-mail: jason.brunson@ufl.edu

and V are overlaid, the scalar product of the i^{th} row marker and the j^{th} column marker approximates x_{ij} ; ditto U and $G = VD$. In general, a biplot overlay of UD^a and VD^b is called *balanced* when $a + b = 1$, so that row–column inner products approximate measurements. The inertia-laden matrices F and G are said to contain *principal coordinates* for the rows and columns of X , while U and V contain *standard coordinates* of unit magnitude. When a matrix factor is represented in principal coordinates, the graphic approximates its geometry in the standard coordinate space of the other factor: In PCA, distances between the markers of F approximate those between the rows of X in the space of measured variables, while inner products and cosines between the markers of G approximate covariances and correlations between the columns of X across the case measurements. In CA, distances between markers of F and of G approximate χ^2 distances between row and column profiles in the count table.

Greenacre [3] uses biplots to surface the mathematical unity of these techniques. However, this unity is rarely reflected in software packages. Implementations of PCA, CA, and several other techniques in `stats` and `MASS`, for example, share no source code beyond `base::svd()`, accept dissimilar arguments for equivalent parameters (e.g. the rank r), and produce output classes with few shared methods. My attempt to harmonize these tools through `ordr`, a tidyverse-inspired API that foregrounds this common theory, is the impetus for the remaining sections.

3 TENSIONS BETWEEN GDA AND GoG

Point clouds are more homogeneous and rigid mathematical abstractions than the case–variable tables for which most statistical graphics are designed. Most conspicuously, the coordinates must be equally scaled to faithfully represent geometric relationships among the markers. This need has not, to my knowledge, received focused attention in the GoG space. For example, while Wilkinson showcases such graphics as Voronoi polygons and sun glyphs that rely for their interpretations on a unit aspect ratio, he only briefly mentions that this choice should be informed by context. On the implementation side, `ggplot2` provides the coordinatization `ggproto::CoordFixed` with default `ratio = 1`, but its invocation is the responsibility of the user. This mismatch between domain and convention has enabled data analysts to publish biplots that misrepresent the geometry of their data.

Some default and hardcoded settings of `ggplot2` are also at odds with best practices in GDA visualization: The artificial coordinates used as scaffolding rarely warrant labels, let alone tick marks or grid lines; instead, Gower et al. [2] recommend that variable axes be incorporated as graphical elements within the plot window. Like coordinate axes, these variable axes must also be calibrated according to any transformations applied to the input data (as in PCA and CA), which requires that the families and parameters of these transformations be made available to the geometric constructor.

Finally, as imbalanced biplots are strongly discouraged, rigorous graphical software should inhibit such choices of the exponents a and b . Biplot-specific programs usually hide or protect this choice from the user: `stats::biplot.princomp()` passes the argument `scale` to b and $1 - \text{scale}$ to a , while `biplotEZ` provides the character-valued argument `CA(variant)` and the logical-valued argument `PCA(correlation.biplot)` that specify row–principal, column–principal, or symmetric (neither–principal) overlays. Because `ggplot2` and similar GoG implementations generally determine plot layers from one data set at a time, this balance is again the responsibility of the user.

4 CONTRIBUTIONS OF GDA TO GoG

The aforementioned tensions should be taken not as intrinsic to the grammatical approach but as an opportunity to advance theory and practice. The `ggplot2` extension community in particular maintains

a tradition of embracing such challenges. I wish to highlight three contributions I believe GDA has to make to GoG, with preliminary implementations in `ordr` and its spinoff `gggda`.

First, it is often of interest in GDA to visually qualify how well the geometry in the first artificial dimensions (those used to plot) captures that of the full space.¹ One approach is to assemble a minimal spanning tree in high dimensions and overlay its projection [4]. Such layers require that the statistical transformation object have access to the full point cloud—inconsistent if not impossible in GoG as originally conceived. Drawing from a tradition of set-valued aesthetics tracing back to `ggmosaic`, `gggda` provides a mechanism to pass coordinate data of arbitrary dimension for statistical transformation before projecting for geometric construction.

Second, the unadjusted overlay of two marker sets can yield messy biplots. A common PCA biplot adjustment offsets calibrated variable axes from the dense core of case markers, which clarifies projections from select markers onto select axes (yielding inner products). To try and err at acceptable offsets is impractical; serviceable algorithms exist but require at least summary information about secondary data—the cases—to compute graphical elements for primary data—the variables. `gggda` implements a family of *referential stats*, built on the prototype `StatRef`, that take a reference data frame as an additional argument. This trick couples with a feature of `ordr` that preserves both matrix factors of an SVD-based technique so that each by default produces layers with the other as referent.

Finally, to evoke the common scale of the artificial axes and reveal differences in the data ranges between them, biplot windows are often made square, rather than truncated in each dimension to the data range. `gggda` provides a coordinatization `ggproto::CoordRect` to achieve this for any desired ratio of window dimensions. This in turn supports the in-development paneling `ggproto::FacetOrd`, which yields a pairs plot for artificial coordinates that follows the rigid requirements above while permitting both fixed and varying panel sizes. These conventions could be used to plot any set of real-valued data but to my knowledge have not previously been introduced.

5 A CASE FOR DOMAIN-CONSTRAINED GRAMMAR

The case studies above lead me to argue that GoG, in theory as well as in practice, should take a deliberate approach to the constraint sets on statistical graphics associated with specific domains. `ggplot2` itself handles Simple Features data delicately, for example, by transforming segments between geographic coordinates into approximate meridians and tucking the coordinatization constructor `coord_sf()` into the geometric element constructor `geom_sf()`; but users may override these behaviors. At an opposite extreme, `gggraph` and `ggtern` are designed to operate entirely without native `ggplot2` constructors. I suggest that a middle ground is required to satisfy the needs of users in, and especially those straddling, diverse specialties: From each specialty should be derived a *graphical dialect*, constituting both permitted objects and constraints on their combination. Such dialects should be based on standards of good practice and formalized mathematically so that they may be unambiguously implemented. At a low level, an implementation should then declare and adopt a domain-specific dialect, for a single plot (perhaps by using a special initializer) or for a section of code (perhaps by setting a global option), that enables features and imposes constraints (perhaps by code completions and suggestions) during plot assembly. At a high level, a family of well-defined dialects should help organize a grand unified GoG at an important meso-scale between its basic objects and its space of all feasible plot types.

¹This is quantified as a ratio of partial to total variances, called *adequacy*.

```

1 pal <- palette.colors(palette = "Okabe-Ito")
2 Lock5withR::Cereal |>
3   ordinate(
4     cols = Calories:Protein,
5     model = ~ prcomp(., scale. = TRUE)
6   ) |>
7   ggbiplot() +
8   geom_rows_point(size = 1.5, color = pal[6]) +
9   geom_rows_text_repel(
10     aes(label = Name),
11     size = 3, color = pal[6]
12   ) +
13   geom_unit_circle(color = pal[7]) +
14   geom_cols_axis(
15     aes(label = name, center = center, scale = scale),
16     size = 4, num = 10, label_dodge = .02,
17     tick_length = .01, text_dodge = .012
18   ) +
19   geom_cols_vector(
20     arrow = arrow(type = "closed"), linewidth = .75,
21     color = pal[7], vector_labels = FALSE
22   ) +
23   stat_rows_projection(
24     subset = 23, ref_subset = c(2, 4, 6),
25     color = pal[4],
26     linetype = "dashed", linewidth = .75
27   ) +
28   coord_rect() +
29   theme_biplot()

```



Figure 1: Row-principal PCA biplot of cereal nutrition data. Packages `grid`, `dplyr`, `ordr`, and `gggda` must be attached. Lines 3–6: The ordination helper uses the default `center = TRUE` from `prcomp` and knows the balanced inertia ($a = 1, b = 0$) of its returned object. Line 15: Aesthetics `center` and `scale` are used to calibrate the axes. Lines 13 and 19–22: Variable markers in standard coordinates are projected from unit vectors onto the 2-D window; shorter lengths indicate worse approximations. Lines 23–27: Referential stat for orthogonal projections from a case marker to an axis approximately recover values in the table; relies on `LayerRef`. Line 28: Subclass of `CoordFixed` that fixes aspect ratios for both coordinate axes, essential for interpreting the above, and window dimensions, which visually signals the geometry. The use of `pal[*]` throughout and of integers rather than names to specify `subset` and `ref_subset` illustrate current limitations.

REFERENCES

- [1] K. R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, Dec. 1971. doi: 10.2307/2334381 1
- [2] J. C. Gower, S. Lubbe, and N. J. Le Roux. *Understanding Biplots*. John Wiley & Sons, Ltd, Jan. 2011. 2
- [3] M. J. Greenacre. *Biplots in Practice*. Fundación BBVA, Bilbao, 2010. 2
- [4] I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2 ed., 2002. doi: 10.1007/b98835 2
- [5] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, NY, 2009. doi: 10.1007/978-0-387-98141-3 1
- [6] L. Wilkinson. *The Grammar of Graphics*. Statistics and Computing. Springer-Verlag, New York, 2 ed., 2005. doi: 10.1007/0-387-28695-0 1