The `tikzviolinplots` package

Pedro Callil-Soares

January 29, 2023

Contents

Abstract

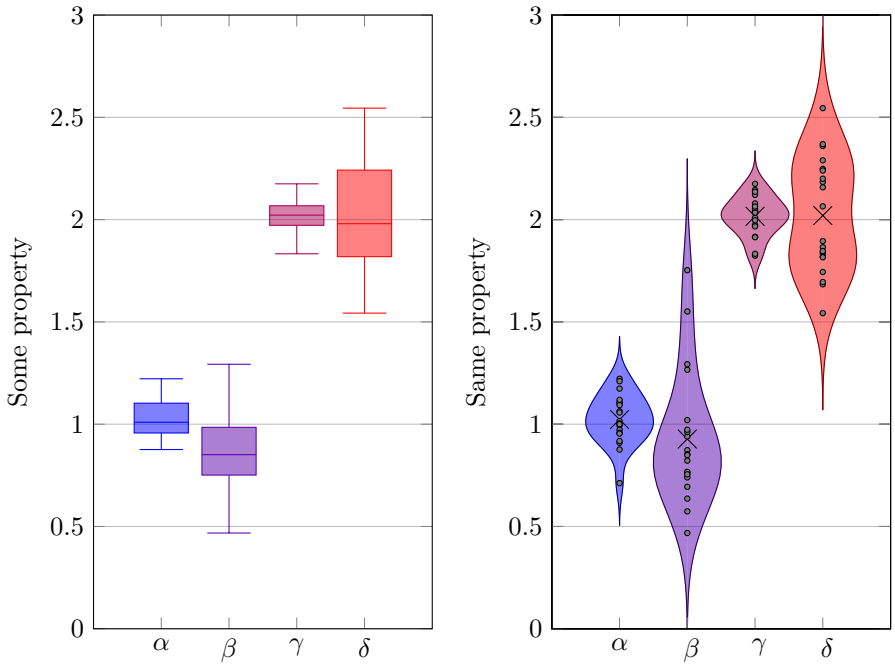
The package provides commands for violin plot creation and the kernel density estimations required.

1 Introduction

This package, through the use of the package `pgfplots`, allows the creation of violin plots in LaTeX. Violin plots are similar to boxplots, but instead of a box signalling the average and quartiles, a kernel density estimator is plotted, as in equation 1, in which the function k (the kernel) is a probability distribution, the positive number h (the bandwidth) is a smoothing factor and n is the sample size.

$$\mathrm{KDE}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right) \tag{1}$$

A comparison between the two plots, showcasing its similarities, is shown in figures 1. The violin plot in figure 1b assumes normal data, and the bandwidth (smoothing factor h in equation 1) is defined accordingly.



(a) Box plot (b) Violin plot

Figure 1: Box and violin plot examples

2 Usage

To plot a violin plot with the commands provided, one must, inside a `tikzpicture` environment, set the general options to all plots and insert each individual dataset.

To set the general options, the command `\violinsetoptions` is provided, and must be invoked before plotting the datasets. This should be done with the commands `\violinplot` or `\violinplotwholefile`.

2.1 General options: `\violinsetoptions`

The command `\violinsetoptions` takes two arguments, an optional argument with package-specific options and a mandatory argument with options to be passed to `pgfplots`.

```
\violinsetoptions[<package-specific options>]%
        {<pgfplots general options>}
```

2.1.1 Package-specific options

There are five options specific to the package: `scaled`, `data points`, `averages`, `no mirror` and `reverse axis`, controlling how and which information in the datasets should be presented.

The option `scaled` controls if all plots in the graph have the same area or same width. If passed, the kernel distribution estimations will be scaled to the same width, as shown in figure 2; otherwise, the plots will present the same area, as in figure 3.

The option `data points`, if passed, will show, along with the violin plots, the distribution of points in the data set, as shown in figure 2.

If the option `averages` is passed, the average of the data set elements is shown, as in figure 3.

The plots are mirrored by default; however, passing the option `no mirror` will show only half the plot, as shown in figure 3.

Finally, to "transpose" the plots (*i.e.* show the distributions as functions of the abcissa, as in figure 3, and not as functions of ordinate, as in figure 2), one might use the option `reverse axis`.

2.1.2 Plot limits and other `pgfplots` options

The minima and maxima of the plot axes must be set in the second (and first mandatory) argument to the command, and should follow `pgfplots` syntax. For instance, to set the minimum and maximum of the x-axis to -3 and 6, and of the y-axis to 2.5 and 7, one might use:

```
\violinsetoptions[<package-specific options>]%
        {xmin=-3, xmax=6, ymin=2.5, xmin=7,%
        <pgfplots general options>}
```

Other `pgfplots` expressions such as title or axes labels may be set in the same way in this argument.

2.2 Options for each data set: `\violinplot`

If the data sets are not very similar and/or advanced customizations are desired, `\violinplot` should be used to plot each data set individually. This command takes one mandatory argument, and a list of options:

```
\violinplot[%
        <option>=<value>
]{filename}
```

The filename (mandatory argument) must be a path to a file storing the data as space-separated columns. The optional argument is a list of options, including:

- `col sep`: Column separation character in filename. Defaults to `comma`, and can accept options `space`, `tab`, `comma`, `colon`, `semicolon`, `braces`, `&` and `ampersand`.
- `index`: Necessary option, is the name of the column with the data to be plotted.
- `kernel`: The function to be used for the kernel density estimation; available values are `gaussian` (default), `logistic`, `parabolic`, `uniform` and `triangular`.
- `bandwidth`: Smoothing parameter for the kernel density estimation; defaults to h in equation 2, which assumes gaussian distribution.

$$h = \sqrt[5]{\frac{4 \times \text{stddev}\left(x_1, x_2, \ldots, x_n\right)^5}{3n}} \tag{2}$$

- `samples`: Number of points in which the kernel density estimation will be calculated and plotted. Defaults to 50.
- `relative position`: position of the plot in the axis with the labels for each dataset. Defaults to 0.
- `color`: Color of the plot, border and interior. Defaults to black. The interior is colored with 0.5 opacity.
- `label`: Label of the plot, positioned at the point set by `relative position`.
- `invert`: If `true`, plot will be drawn in opposite side. Useful for comparing two parts of a data set, by plotting each group side by side.
- Options for average and data points:
  - `average mark`, `dataset mark`: Mark used for averages and points of the data set (defaults to "`x`" and "`*`", respectively).
  - `average size`, `dataset size`: Size of each mark (defaults to 3pt for averages and 1pt for other points).
  - `average color`, `dataset color`: Color of each mark's border. Defaults to black in both cases.
  - `average opacity`, `dataset opacity`: Border opacity. Defaults to 1.0 in both cases.
  - `average fill`, `dataset fill`: Color of each mark's interior. Defaults to black for the points of the data set and white for the averages.
  - `average fill opacity`, `dataset fill opacity`: Opacity of the marks. Defaults to 0.5 for averages and 0.2 for the other points.

2.3 Simplified interface: `\violinplotwholefile`

If the data sets are similar and no customizations are required, one might use the command `\violinplotwholefile`.

```
\violinplotwholefile[%
        <option>=<value>
]{filename}
```

This command will calculate and plot all columns named. The available options are:

- `primary color`: Primary color utilised. A gradient is built from the secondary color to it. Defaults to black.
- `secondary color`: Secondary color utilized. A gradient is built from it to the primary color. Defaults to white.
- `indexes`: List of columns to be plotted.
- `spacing`: Distance between plots. Defaults to 1.0.
- `labels`: Labels of the data to be plotted.

Besides, the options available to `\violinplots` are also available for this command; for instance, one might set the kernel to be utilized as the `uniform` kernel through the option `kernel=uniform`. Of course, `relative position` is not an option here.

3 Examples

To show the usage of the package, we plotted several data sets two times, in figures 2 and 3, using either the simplified or the complete interface.

The data are in the file `example.dat`, shown below:

```
A          B          C          D          E
0.3        -2.1       3.50       2.89       1.00
0.41       -1.9       3.55       2.88       1.06
0.45       -1.5       3.55       3.13       1.00
0.46       -1.3       3.60       2.69       1.20
0.46       -1.3       3.60       2.78       1.00
0.46       -1.27      3.60       2.83       1.35
0.47       -1.26      3.65       3.08       1.00
0.47       -1.26      3.65       3.08       1.53
0.48       -1.24      3.65       2.73       1.00
0.51       -1.2       3.65       3.08       1.73
0.57       -1.13      3.65       3.24       1.00
```

```
2.3     -1.02   3.70    3.10    1.95
2.41    -0.9    3.70    2.98    1.00
2.46    -0.2    3.70    2.98    2.21
2.47    0.0     3.75    3.04    1.00
2.48    0.1     3.80    3.24    2.49
2.51    0.3     3.85    3.16    1.00
2.57    0.5     3.85    3.30    3.04
```

3.1 Simplified Interface

First we will plot this data using the simplified interface, resulting in figure 2.

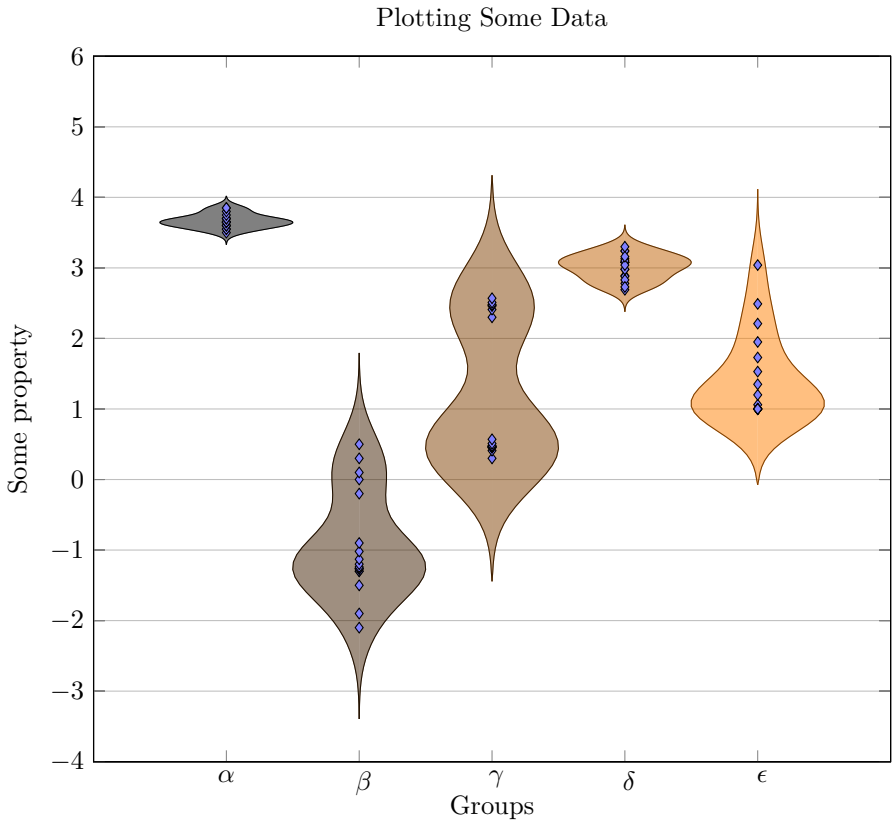


Figure 2: Violin plot — vertical example

Code for figure 2, using `\violinplotwholefile`:

```
\begin{tikzpicture}
    \violinsetoptions[
        data points,
        scaled,
    ]{
        xmin=0,xmax=6,
        ymin=-4,ymax=6,
        title={Plotting Some Data},
        xlabel={Groups},
        ylabel={Some property},
        xlabel style={
            yshift = {-2*height("a")}
        },
        ymajorgrids=true,
    }
    \violinplotwholefile[%
        primary color=orange,
        secondary color=black,
        batch indexes={C,B,A,D,E},
        batch spacing=1.0,
        batch labels={%
            $\alpha$,
            $\beta$,
            $\gamma$,
            $\delta$,
            $\epsilon$
        },
        col sep=tab,
        dataset size=2pt,
        dataset mark=diamond*,
        dataset fill=blue!50!white,
        dataset fill opacity=1.0,
    ]{example.dat}
\end{tikzpicture}
```

3.2 Complete Interface

Then the data will be plotted using the complete interface, resulting in figure 3.

Code for figure 3, using `\violinplot`:

```
\begin{tikzpicture}
    \violinsetoptions[
```

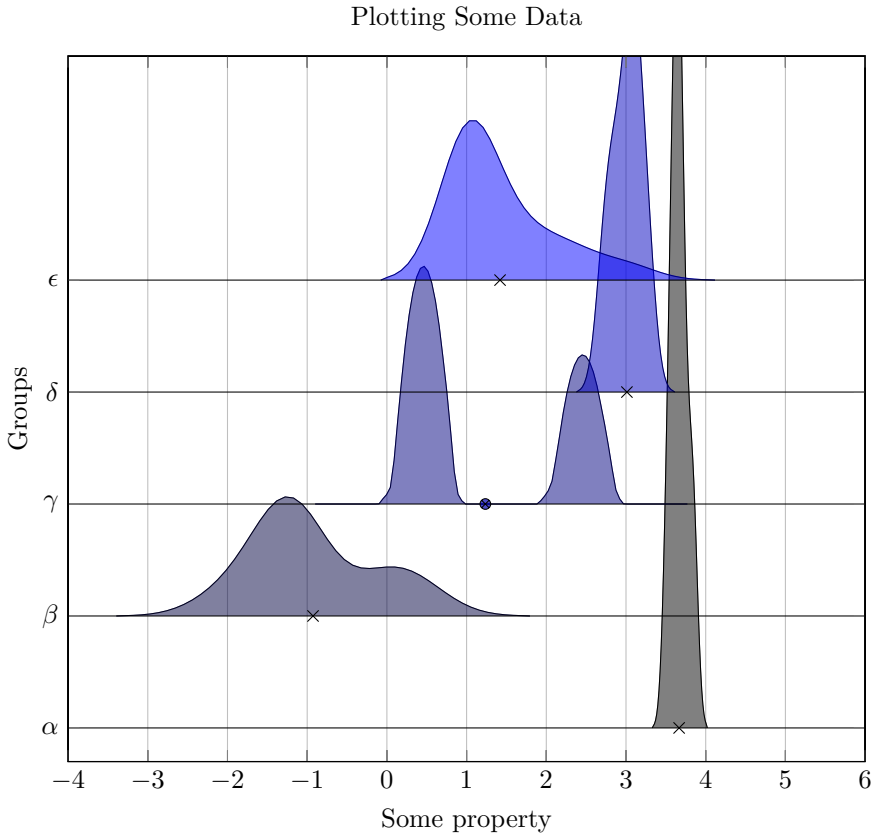


Figure 3: Violin plot — horizontal example

```
        no mirror,
        averages,
        reverse axis,
    ]{%
        xmin=-4,xmax=6,
        ymin=0.7,ymax=7,
        title={Plotting Some Data},
        ylabel={Groups},
        xlabel={Some property},
        ylabel style={
            yshift = {2*width("a")}
        },
        xmajorgrids=true,
    }
    \violinplot[%
```

```
        index=C,
        relative position=1,
        color=blue!0!black,
        label={$\alpha$},
        col sep=tab,
    ]{example.dat}
    \violinplot[%
        index=B,
        relative position=2,
        color=blue!25!black,
        label={$\beta$},
        col sep=tab,
    ]{example.dat}
    \violinplot[%
        index=A,
        relative position=3,
        color=blue!50!black,
        label={$\gamma$},
        col sep=tab,
        kernel=parabolic,
        samples=100,
        bandwidth=0.4,
        average mark=otimes*,
        average size=2pt,
        average fill=blue!70!black,
        average fill opacity=0.7,
    ]{example.dat};
    \violinplot[%
        index=D,
        relative position=4,
        color=blue!75!black,
        label={$\delta$},
        col sep=tab,
    ]{example.dat}
    \violinplot[%
        index=E,
        relative position=5,
        color=blue!100!black,
        label={$\epsilon$},
        col sep=tab,
    ]{example.dat}
\end{tikzpicture}
```

3.3 Drawings and Annotations

As everything happens inside a `tikzpicture` environment, drawings and annotations are straightforward, as seen in figure 4.

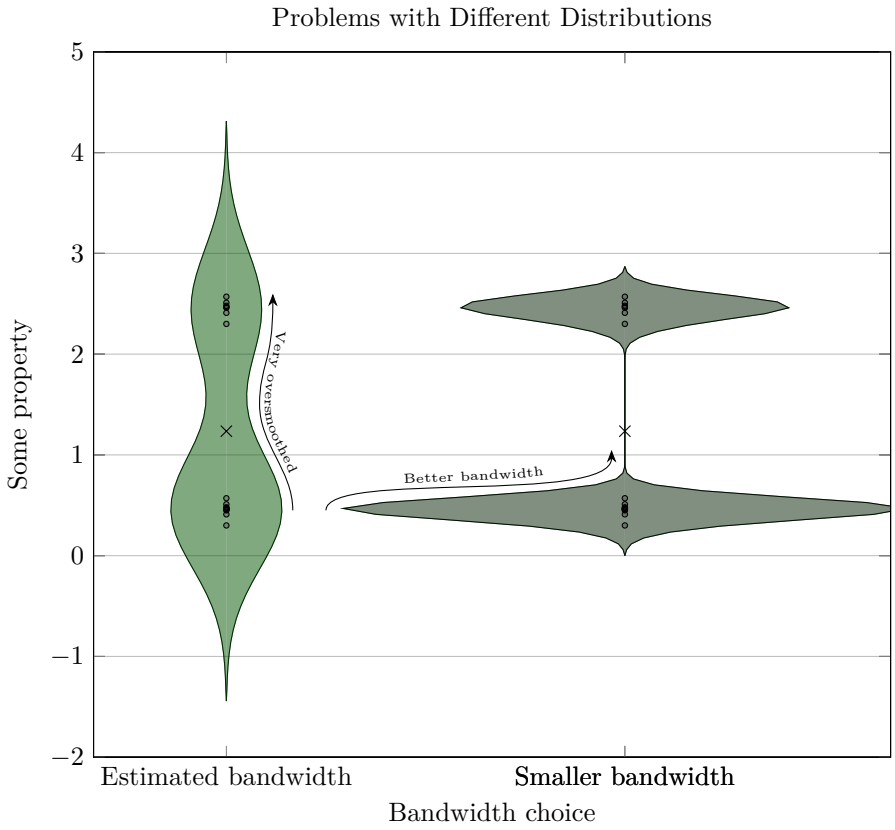


Figure 4: Annotations in a plot

Figure 4 is compiled from the code excerpt below.

```
\begin{tikzpicture}
    \violinsetoptions[
        averages,
        data points,
    ]{
        xmin=0,xmax=6,
        ymin=-2,ymax=5,
        xlabel = {Bandwidth choice},
        ylabel = {Some property},
```

```
        title={Problems with Different %
            Distributions},
        xlabel style={
            yshift = {-3*height("a")}
        },
        ymajorgrids=true,
    }
    \violinplot[%
        index=A,
        relative position=1,
        color=green!33!black,
        label={Estimated bandwidth},
        col sep=tab,
    ]{example.dat}
    \violinplot[%
        index=A,
        bandwidth=0.1,
        relative position=4,
        color=green!12!black,
        label={Smaller bandwidth},
        col sep=tab,
    ]{example.dat}
    \begin{axis}[
        xmin=0,xmax=6,
        ymin=-2,ymax=5,
    ]
    \draw[-{Stealth}] (axis cs: 1.75,0.45) %
    .. controls (axis cs:1.75,0.85) and %
    (axis cs:3.9,0.5) .. (axis cs: 3.9,1.05);
    \draw[decoration={raise=2pt,text along path,%
    text={ Better bandwidth}, %
    text align={center}}, decorate] %
    (axis cs: 1.75,0.45) %
    .. controls (axis cs:1.75,0.85) and %
    (axis cs:3.9,0.5) .. (axis cs: 3.9,1.05);
    \draw[-{Stealth}] %
    (axis cs: 1.5,0.45) .. controls %
    (axis cs:1.5,0.85) and (axis cs:1.25,1.1) %
    .. (axis cs: 1.25,1.5) .. controls %
    (axis cs:1.25,1.9) and (axis cs:1.35,1.9) %
    .. (axis cs:1.35,2.6);
    \draw[decoration={raise=2pt,text along path, %
    text={Very oversmoothed}, %
    text align={center}}, decorate] %
    (axis cs: 1.35,2.6) .. controls %
```

```
    (axis cs:1.35,1.9) and (axis cs:1.25,1.9) %
    .. (axis cs: 1.25,1.5) .. %
    controls (axis cs:1.25,1.1) and %
    (axis cs:1.5,0.85) .. (axis cs:1.5,0.45);
    \end{axis}
\end{tikzpicture}
```

3.4 Standalone Kernel Density Estimation

Plotting only one dataset and choosing adequate plot limits, one may obtain a simple representation of a kernel density estimation. This can be seen in figure 5, obtained from the code excerpt below.

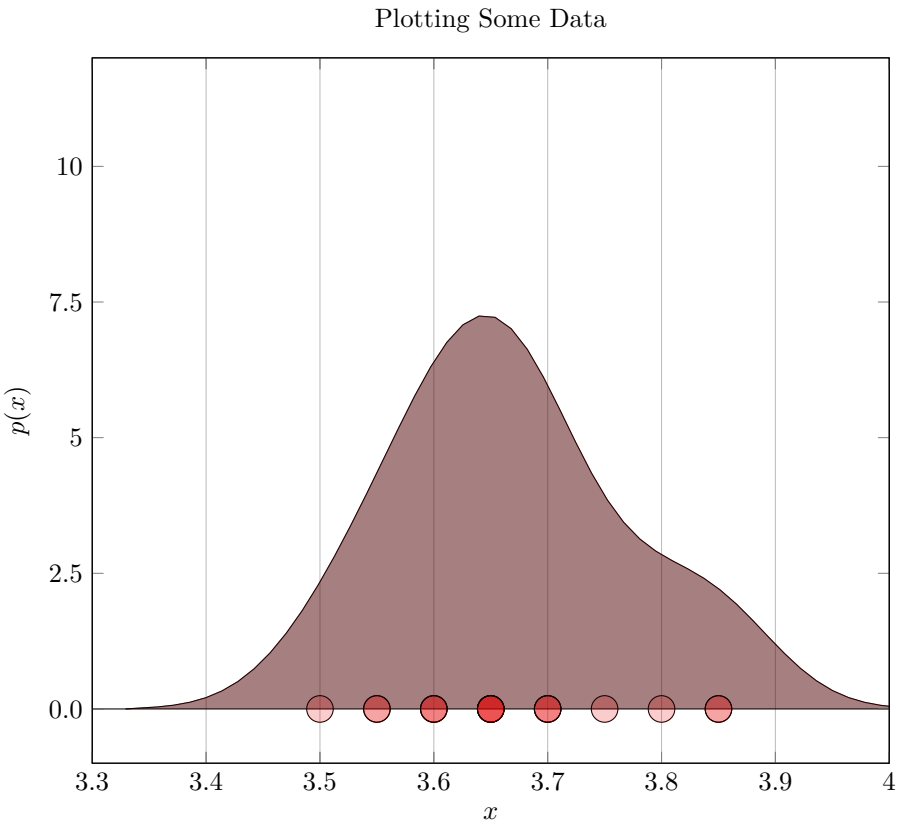


Figure 5: Standalone kernel density estimation example

```
\begin{tikzpicture}
    \violinsetoptions[
        no mirror,
        reverse axis,
        data points,
    ]{%
        xmin=3.3,xmax=4,
        ymin=-1,ymax=12,
        title={Plotting Some Data},
        ylabel={$p(x)$},
        xlabel={$x$},
        xmajorgrids=true,
        ymajorticks=true,
        ytick={0.0,2.5,5,7.5,10},
        yticklabels={0.0,2.5,5,7.5,10},
    }
    \violinplot[%
        index=C,
        relative position=0,
        color=red!30!black,
        label={},
        col sep=tab,
        dataset size=5pt,
        dataset color=red!10!black,
        dataset mark=*,
        dataset fill=red!90!black,
        dataset fill opacity=0.2,
    ]{example.dat}
\end{tikzpicture}
```

3.5 Asymmetrical Violin Plots — Real World Data

Using the key `invert=true` in the options of `\violinplot` one can plot two different sets of data side by side, in an asymmetrical violin plot. This can be seen in figure 6, that exhibits the male and female life expectancy at birth, in 2019, for several countries, segregated in the WHO regions.[1].

The code for figure 6 is avaliable below:

```
\begin{tikzpicture}
    \violinsetoptions[
```

[1] https://www.who.int/data/gho/publications/world-health-statistics.

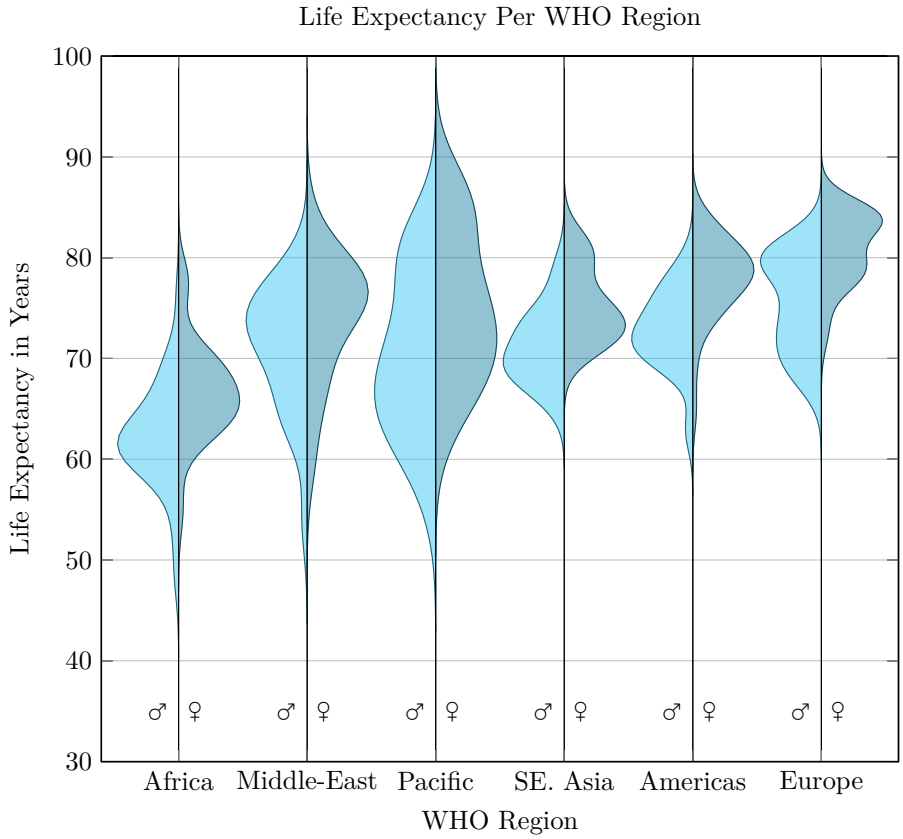


Figure 6: Example of asymmetrical plot

```
        no mirror,
        scaled,
    ]{%
        xmin=0,xmax=14,
        ymin=30,ymax=100,
        title={Life Expectancy Per WHO Region},
        xlabel={WHO Region},
        ylabel={Life Expectancy in Years},
        xlabel style={
            yshift = {-3*width("a")}
        },
        ymajorgrids=true,
    }
    \violinplot[%
        index=LifeExpectancyAtBirthMale,
```

```
        relative position=2,
        color=malecolor,
        label={AFR},
        col sep=comma,
        invert=true
    ]{AFR.csv}
    \violinplot[%
        index=LifeExpectancyAtBirthFemale,
        relative position=2,
        color=femalecolor,
        label={},
        col sep=comma,
    ]{AFR.csv}
    \violinplot[%
        index=LifeExpectancyAtBirthMale,
        relative position=4,
        color=malecolor,
        label={EMR},
        col sep=comma,
        invert=true
    ]{EMR.csv}
    \violinplot[%
        index=LifeExpectancyAtBirthFemale,
        relative position=4,
        color=femalecolor,
        label={},
        col sep=comma,
    ]{EMR.csv}
    \violinplot[%
        index=LifeExpectancyAtBirthMale,
        relative position=6,
        color=malecolor,
        label={WPR},
        col sep=comma,
        invert=true
    ]{WPR.csv}
    \violinplot[%
        index=LifeExpectancyAtBirthFemale,
        relative position=6,
        color=femalecolor,
        label={},
        col sep=comma,
    ]{WPR.csv};
    \violinplot[%
        index=LifeExpectancyAtBirthMale,
```

```
        relative position=8,
        color=malecolor,
        label={SEAR},
        col sep=comma,
        invert=true
    ]{SEAR.csv}
    \violinplot[%
        index=LifeExpectancyAtBirthFemale,
        relative position=8,
        color=femalecolor,
        label={},
        col sep=comma,
    ]{SEAR.csv}
    \violinplot[%
        index=LifeExpectancyAtBirthMale,
        relative position=10,
        color=malecolor,
        label={AMR},
        col sep=comma,
        invert=true
    ]{AMR.csv}
    \violinplot[%
        index=LifeExpectancyAtBirthFemale,
        relative position=10,
        color=femalecolor,
        label={},
        col sep=comma,
    ]{AMR.csv}
    \violinplot[%
        index=LifeExpectancyAtBirthMale,
        relative position=12,
        color=malecolor,
        label={EUR},
        col sep=comma,
        invert=true
    ]{EUR.csv}
    \violinplot[%
        index=LifeExpectancyAtBirthFemale,
        relative position=12,
        color=femalecolor,
        label={},
        col sep=comma,
    ]{EUR.csv}
    \begin{axis}[
        xmin=0,xmax=14,
```

```
        ymin=30,ymax=100,
    ]
        \draw(axis cs:2,35) node[anchor=east] {\male};
        \draw(axis cs:2,35) node[anchor=west] {\female};
        \draw(axis cs:4,35) node[anchor=east] {\male};
        \draw(axis cs:4,35) node[anchor=west] {\female};
        \draw(axis cs:6,35) node[anchor=east] {\male};
        \draw(axis cs:6,35) node[anchor=west] {\female};
        \draw(axis cs:8,35) node[anchor=east] {\male};
        \draw(axis cs:8,35) node[anchor=west] {\female};
        \draw(axis cs:10,35) node[anchor=east] {\male};
        \draw(axis cs:10,35) node[anchor=west] {\female};
        \draw(axis cs:12,35) node[anchor=east] {\male};
        \draw(axis cs:12,35) node[anchor=west] {\female};
    \end{axis}
\end{tikzpicture}
```

4 Limitations

As the math is handled through TEX, generating the kernel distribution estimations is a slow process; therefore, if possible, the number of samples should not be very large. Also, each value should not, itself, be too large, to avoid "dimension too large" errors. In this case, the data should be scaled and the tick labels manually corrected.

As each violin plot is rendered in a different `axis` environment, the positions of the axes' labels are not, usually, correct. One should set it manually. In the examples, this was accomplished setting the vertical/horizontal shift to twice the height of a letter, in the direction opposite to the appropriate axis. This approach also leads to another problem: using the keys "height" or "width" won't rescale the plot; one might use, before and after the code for the figure, `\pgfplotsset`, as shown below:

```
\pgfplotsset{height=1.75\linewidth}
\begin{tikzpicture}
    \violinsetoptions[
        options for tikzviolinplots
    ]{
        pgfplots options
    }
    \violinplot[%
        options for data points
    ]{filename}
```

```
    ...
\end{tikzpicture}
\pgfplotsset{height=0.9\linewidth}
```

For the same reason, the options `xmin`, `ymin`, `xmax` and `ymax` are required; no automatic placement is performed.