ggdensity: Improved density visualization in R

James Otto, David Kahle

Department of Statistical Science, Baylor University

June 21, 2022

Introduction

- `ggplot2` includes several ways to estimate and visualize densities for uni- and bivariate data
  - Limited by the difficulty of interpreting density height

Introduction

- `ggplot2` includes several ways to estimate and visualize densities for uni- and bivariate data
  - Limited by the difficulty of interpreting density height
- `ggdensity` extends `ggplot2`
  - Interpretable visualizations via highest density regions

Motivating Example

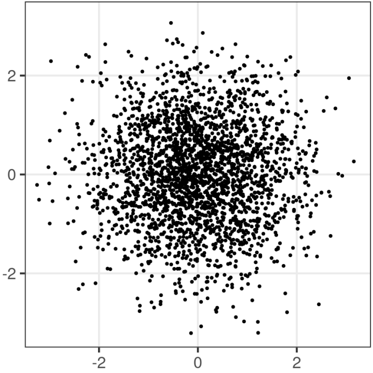


Figure 1: Simulated bivariate standard normal sample ($n = 2500$)

Motivating Example

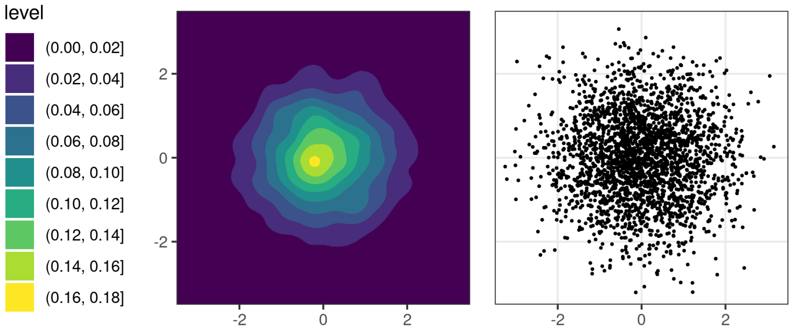


Figure 2: Visualizing density estimate with `geom_density2d_filled`

Motivating Example

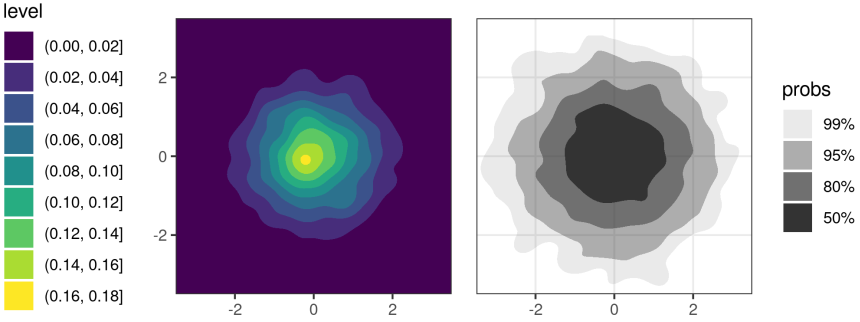


Figure 3: Comparing `geom_density2d_filled` (left) and `geom_hdr` (right)

Highest Density Regions

- Advantages to plotting HDRs instead of arbitrary density contours:
  - Inferentially relevant
  - Interpretable

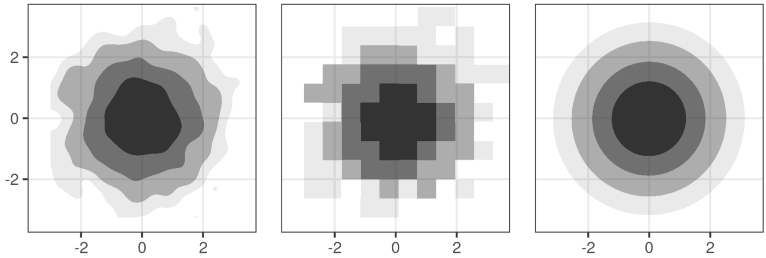
Highest Density Regions

- Advantages to plotting HDRs instead of arbitrary density contours:
  - Inferentially relevant
  - Interpretable
- Estimated HDRs depend on estimated density surface
  - Different estimators $\implies$ different HDRs

geom_hdr

Exploring choices of density estimator

```
df <- tibble(x = rnorm(1000), y = rnorm(1000))

ggplot(df, aes(x, y)) + geom_hdr()
ggplot(df, aes(x, y)) + geom_hdr(method = "histogram")
ggplot(df, aes(x, y)) + geom_hdr(method = "mvnorm")
```



geom_hdr

Exploring choices of density estimator

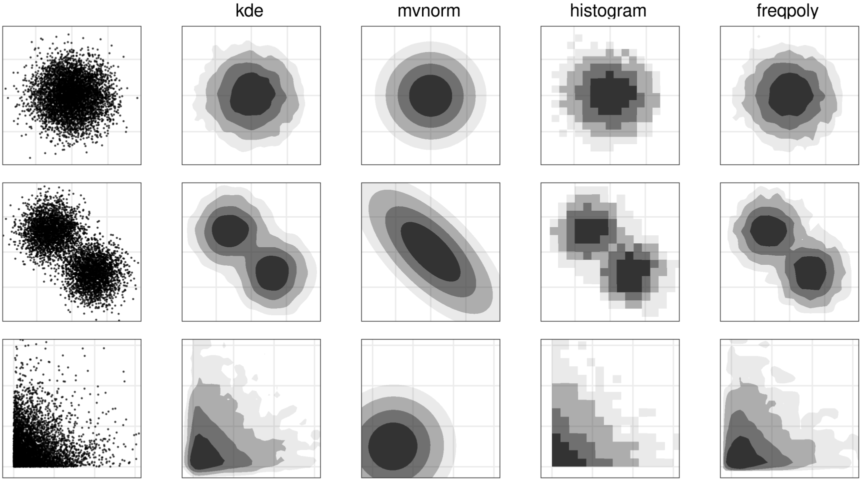
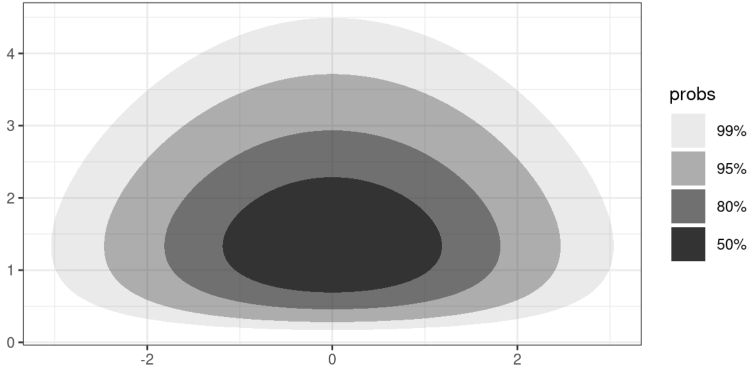


Figure 4: HDRs resulting from different choices of $\hat{f}$

geom_hdr_fun

Plotting HDRs from a known parametric density

```r
f <- function(x, y) dnorm(x) * dgamma(y, 5, 3)

ggplot() +
  geom_hdr_fun(fun = f, xlim = c(-4, 4), ylim = c(0, 5))
```



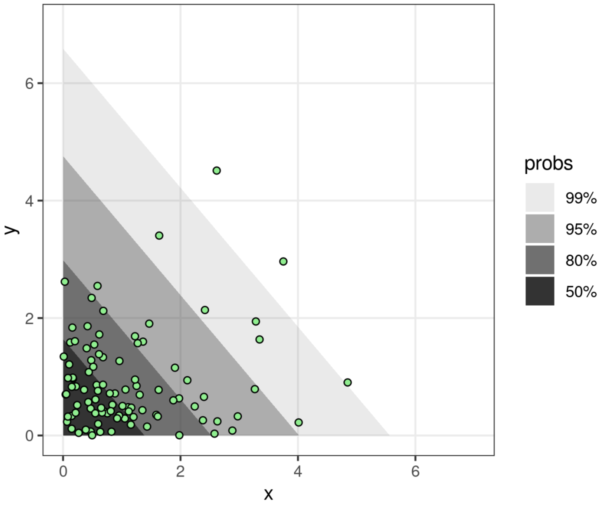
geom_hdr_fun

Plotting HDRs from an estimated parametric density

```
df <- data.frame(x = rexp(100, 1), y = rexp(100, 1))

# pdf for parametric density estimate
f <- \(x, y, lambda) dexp(x, lambda[1]) * dexp(y, lambda[2])

# estimate parameters governing joint pdf
lambda_hat <- apply(df, 2, mean)

# make plot
ggplot(df, aes(x, y)) +
  geom_hdr_fun(fun = f, args = list(lambda = lambda_hat)) +
  geom_point(fill = "lightgreen", shape = 21)
```

geom_hdr_fun

Plotting HDRs from an estimated parametric density



Palmer Penguins

The Palmer penguins data set contains various measurements for three penguin species located in the Palmer Archipelago, Antarctica:

```
## # A tibble: 344 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex
##   <fct>   <fct>           <dbl>         <dbl>            <int>       <int> <fct>
## 1 Chinst~ Dream            49            19.6              212        4300 male
## 2 Gentoo  Biscoe           45.8          14.6              210        4200 fema~
## 3 Adelie  Torge~           39            17.1              191        3050 fema~
## 4 Chinst~ Dream            43.2          16.6              187        2900 fema~
## 5 Gentoo  Biscoe           48.8          16.2              222        6000 male
## 6 Gentoo  Biscoe           49.1          14.8              220        5150 fema~
## 7 Chinst~ Dream            40.9          16.6              187        3200 fema~
## # ... with 337 more rows, and 1 more variable: year <int>
```

Palmer Penguins

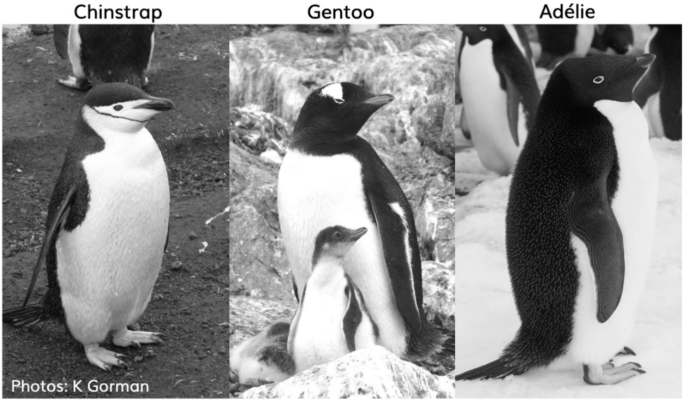


Figure 5: Examples of the three species of penguins

Palmer Penguins

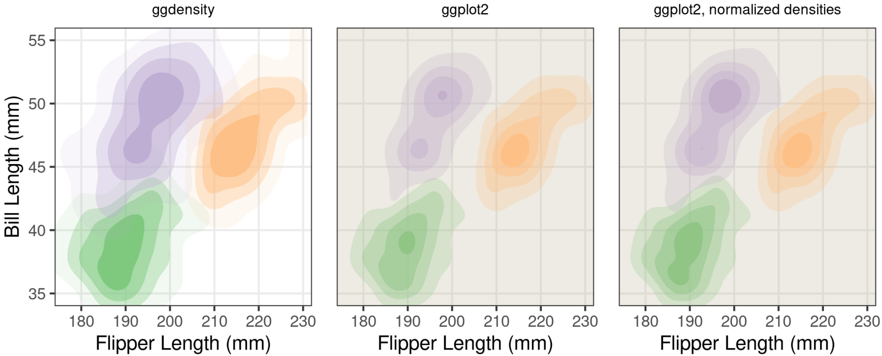


Figure 6: Comparing grouping with Palmer penguins data

Palmer Penguins

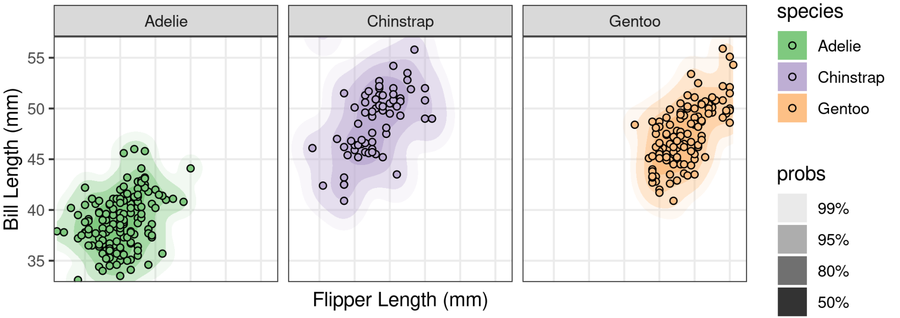
The code to generate the plots in figure 6 showcases another advantage of `ggdensity`:

```
ggplot(penguins, aes(flipper_length_mm, bill_length_mm, fill = species)) +
  geom_hdr(probs = c(.95, .8, .6, .3))

ggplot(penguins, aes(flipper_length_mm, bill_length_mm, fill = species)) +
  geom_density_2d_filled(aes(alpha = after_stat(level)),
                         contour_var = "ndensity", bins = 4)
```

In order to create the plot with `geom_density2d_filled`, the user needs to be aware of several advanced `ggplot2` concepts

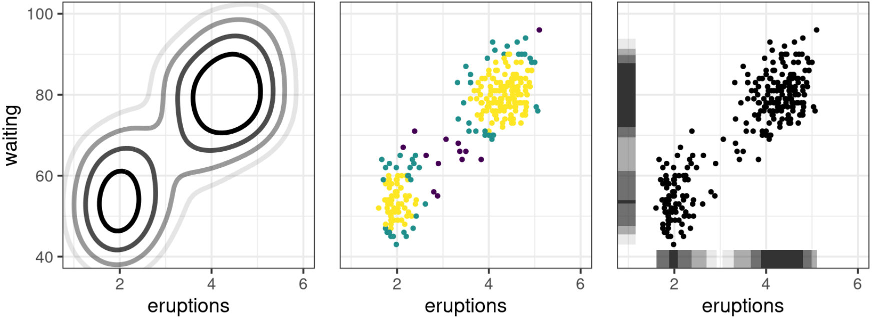
Palmer Penguins

```
ggplot(penguins, aes(flipper_length_mm, bill_length_mm, fill = species)) +
  geom_hdr() +
  geom_point(shape = 21) +
  facet_wrap(vars(species))
```



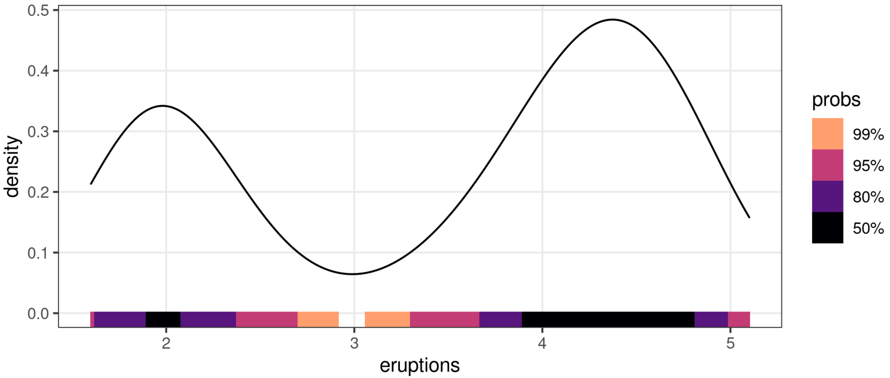
Old Faithful

```r
p <- ggplot(faithful, aes(eruptions, waiting))

p + geom_hdr_lines()
p + geom_hdr_points()
p + geom_hdr_rug()
```



Old Faithful

```
ggplot(faithful, aes(eruptions)) +
  geom_density() +
  geom_hdr_rug(aes(fill = after_stat(probs)), alpha = 1) +
  scale_fill_viridis_d(option = "magma", begin = .8, end = 0)
```



Related Projects

- `hdrcde`
  - Bivariate HDR plots using base graphics
  - Many technical differences

Related Projects

- `hdrcde`
  - Bivariate HDR plots using base graphics
  - Many technical differences
- `gghdr`

Related Projects

- `hdrcde`
  - Bivariate HDR plots using base graphics
  - Many technical differences
- `gghdr`
- `ggdist`

References

Azzalini, A. and A. W. Bowman (1990). "A Look at Some Data on the Old Faithful Geyser". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 39.3, pp. 357–365. ISSN: 00359254, 14679876.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman (2020). *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. R package version 0.1.0.

Hyndman, Rob et al. (Jan. 2021). *hdrcde: Highest Density Regions and Conditional Density Estimation*.

Hyndman, Rob J. (1996). "Computing and Graphing Highest Density Regions". In: *The American Statistician* 50.2, pp. 120–126. ISSN: 00031305.

Kay, Matthew (2022). *ggdist: Visualizations of Distributions and Uncertainty*. R package version 3.1.1. DOI: `10.5281/zenodo.3879620`.

O'Hara-Wild, Mitchell et al. (Feb. 2022). *gghdr: Visualisation of Highest Density Regions in 'ggplot2'*.

Scott, David (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. ISBN: 9781118575536.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4.

Wickham, Hadley et al. (2019). "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43, p. 1686. DOI: `10.21105/joss.01686`.

Wilkinson, Leland (2005). *The Grammar of Graphics (Statistics and Computing)*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387245448.

Thank you!

jamesotto852.github.io

@jamesotto852

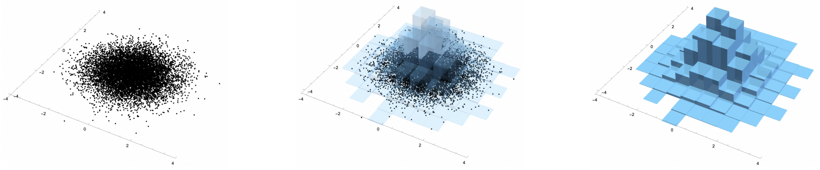
Definition

Let $f(x)$ be the density function of a random variable X. Then the $100(1-\alpha)\%$ highest density region (HDR) is the subset $R(f_\alpha)$ of the sample space of X such that $R(f_\alpha) = \{x : f(x) \geq f_\alpha\}$ where f_α is the largest constant such that $\mathrm{P}\left(X \in R(f_\alpha)\right) \geq 1 - \alpha$.

R. J. Hyndman 1996

Additional Materials

Illustrating the Numerical Integration Method



(a) (b) (c)

Figure 7: Estimating 3-dimensional histogram surface

Additional Materials

Illustrating the Numerical Integration Method

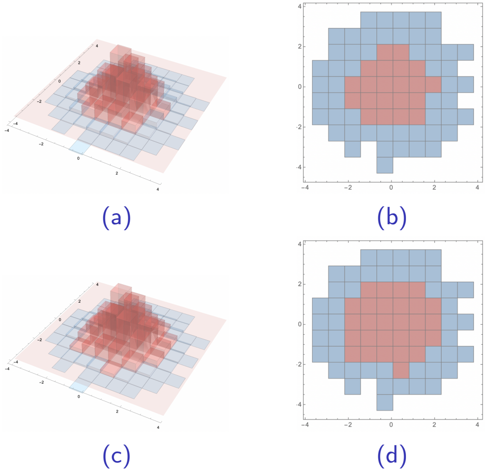


Figure 8: Calculating resulting 75% and 90% HDRs

Additional Materials

Illustrating the Numerical Integration Method

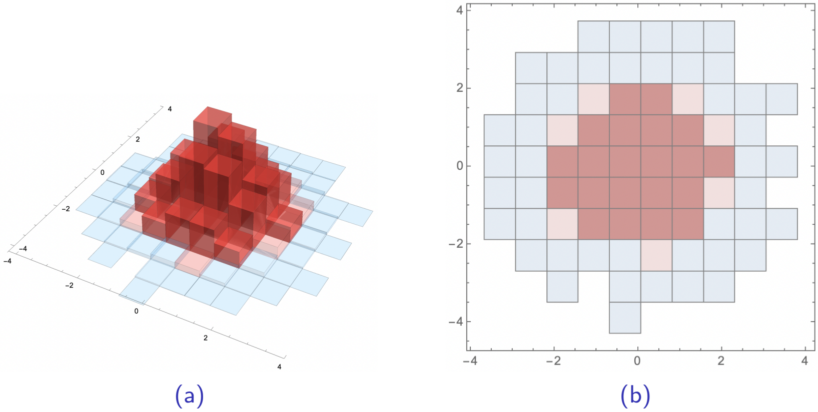


Figure 9: Visualizing 75% and 90% HDRs together