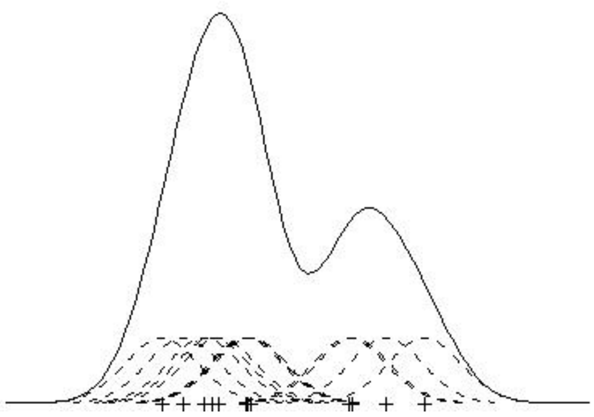
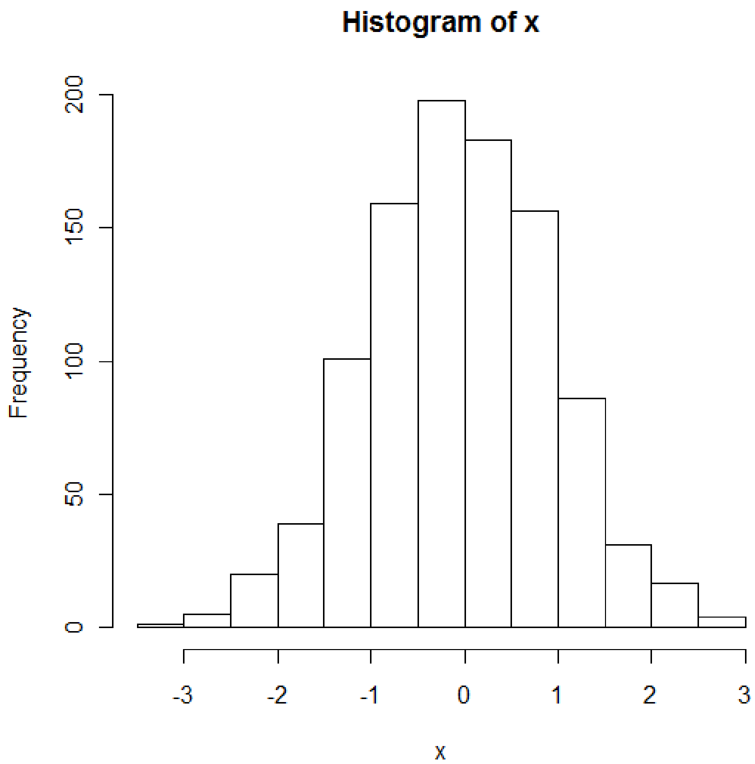
Density Estimation in R

Henry Deng & Hadley Wickham



Density Estimation

- Focus on univariate, nonparametric
- Helps reveal underlying distributions
- Applicable in real-life scenarios
- Utility as intermediate step for other calculations



Motivation

- Over 25 packages in R that contain density estimation functions
  - Fifteen suitable for our specific needs
- Provide how and how well packages worked
- Packages rely on differing mathematical theoretical approaches
- Wanted to evaluate performance among the density estimation functions in the packages
- Benefits standard R users, developers

Procedures

- Identify which packages to study
- Theoretical overview of all packages
  – Reference manuals, articles, books
- Compare calculation speed and accuracy
  – Run tests to evaluate performance
- Summarize findings and investigate other ideas
  – Link theory and performance

Packages Studied

Package	Function	Dimensions	Approach
ASH	ash1	2	ASH
base	density	1	Kernel
ftnonpar	pmden	1	Taut String
GenKern	KernSec	2	Kernel
gss	dssden	≥1	Penalized
MASS	hist	1	Histogram
kerdiest	kde	1	Kernel
KernSmooth	bkde	2	Kernel
ks	kde	6	Kernel
locfit	density.lf	1	Local Likelihood
logspline	dlogspline	1	Penalized
np	npudens	1	Kernel
pendensity	pendensity	1	Penalized
plugdensity	plugin.density	1	Kernel
sm	sm.density	3	Kernel

Theoretical Approach

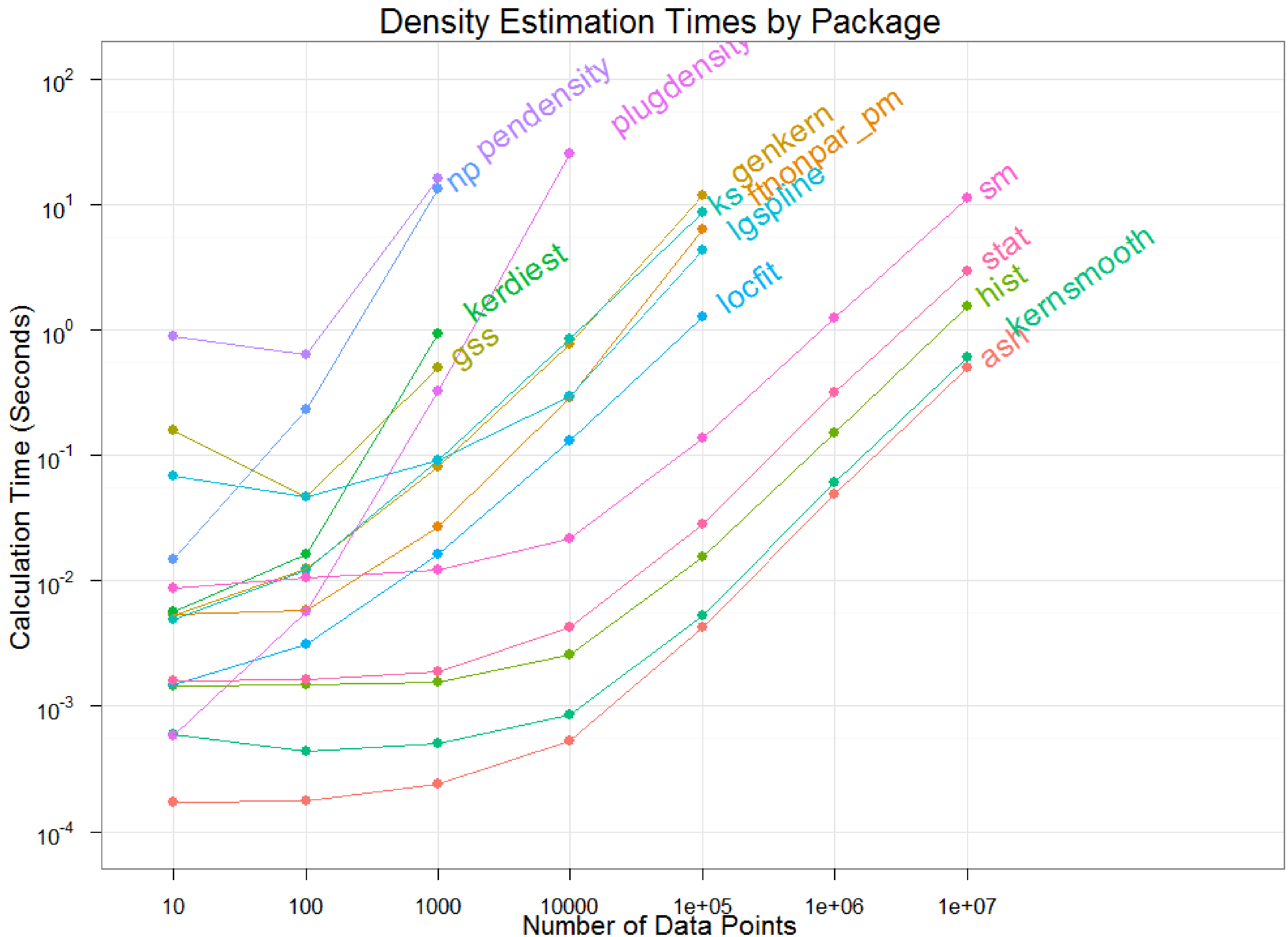
- Methods for Density Estimation
  - Histogram Approach
  - Kernel Density Estimation
  - Other techniques
    - Penalized Methods, Taut Strings, Splines

ASH: $$\widehat{f}(x;m) = \frac{1}{nh} \sum_{|i|<m} w_m(i) v_{k+i}$$

KDE: $$\widehat{f}(x,H) = \frac{1}{n} \sum_{i=1}^{n} K_H(x - x_i)$$

Calculation Speed

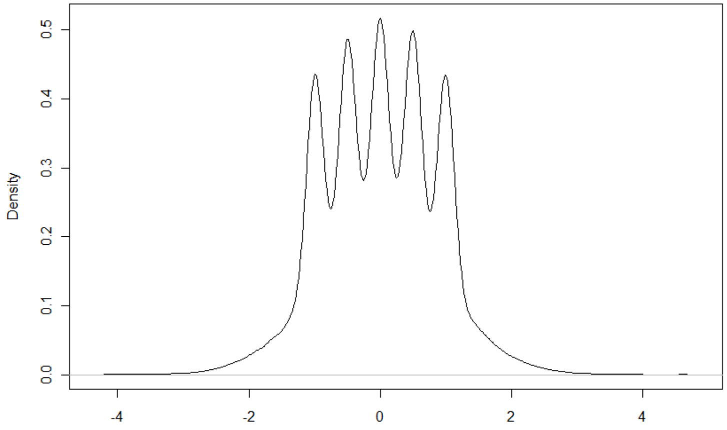
- Procedure
  - Random set of *n* normally distributed points
  - Increasing number of points (*n*)
  - Multiple trials
- Timing
  - Microbenchmark package to record time
  - Measures nanoseconds

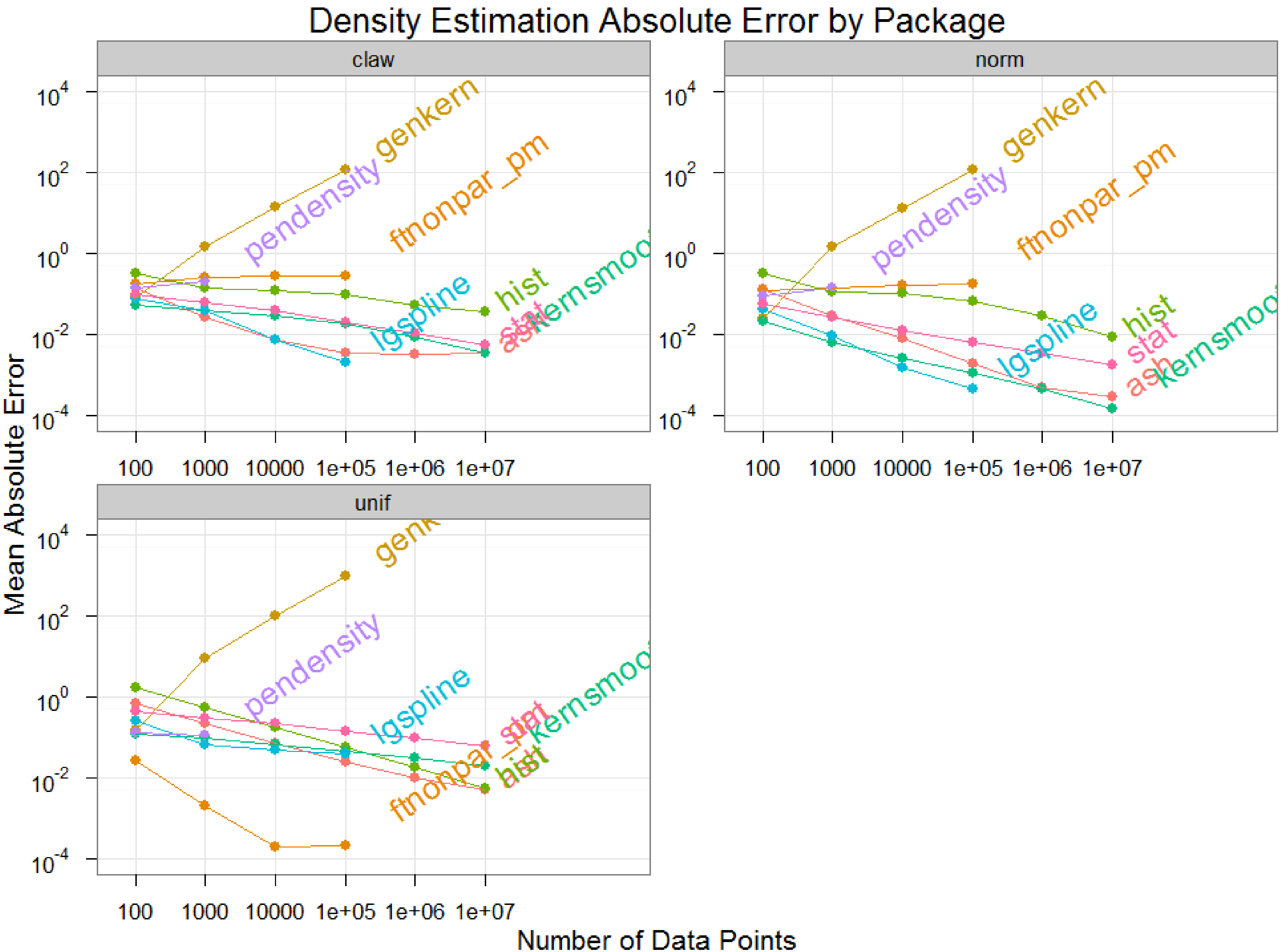


Density Estimation Times by Package
Calculation Time (Seconds)
Number of Data Points
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10
100
1000
10000
1e+05
1e+06
1e+07
np
pendensity
plugdensity
kerdiest
gss
genkern
ftnonpar_pm
ks
lgspline
locfit
sm
stat
hist
kernsmooth
ash

Estimation Accuracy

- Specifications
  - Distribution: uniform, normal, claw
  - Grid density evaluation points by 512
  - Used default parameters: automatic bandwidth selection, etc
  - Increasing number of data points
  - Multiple trials
- Measuring Error
  - Mean Absolute Error
  - Mean Squared Error

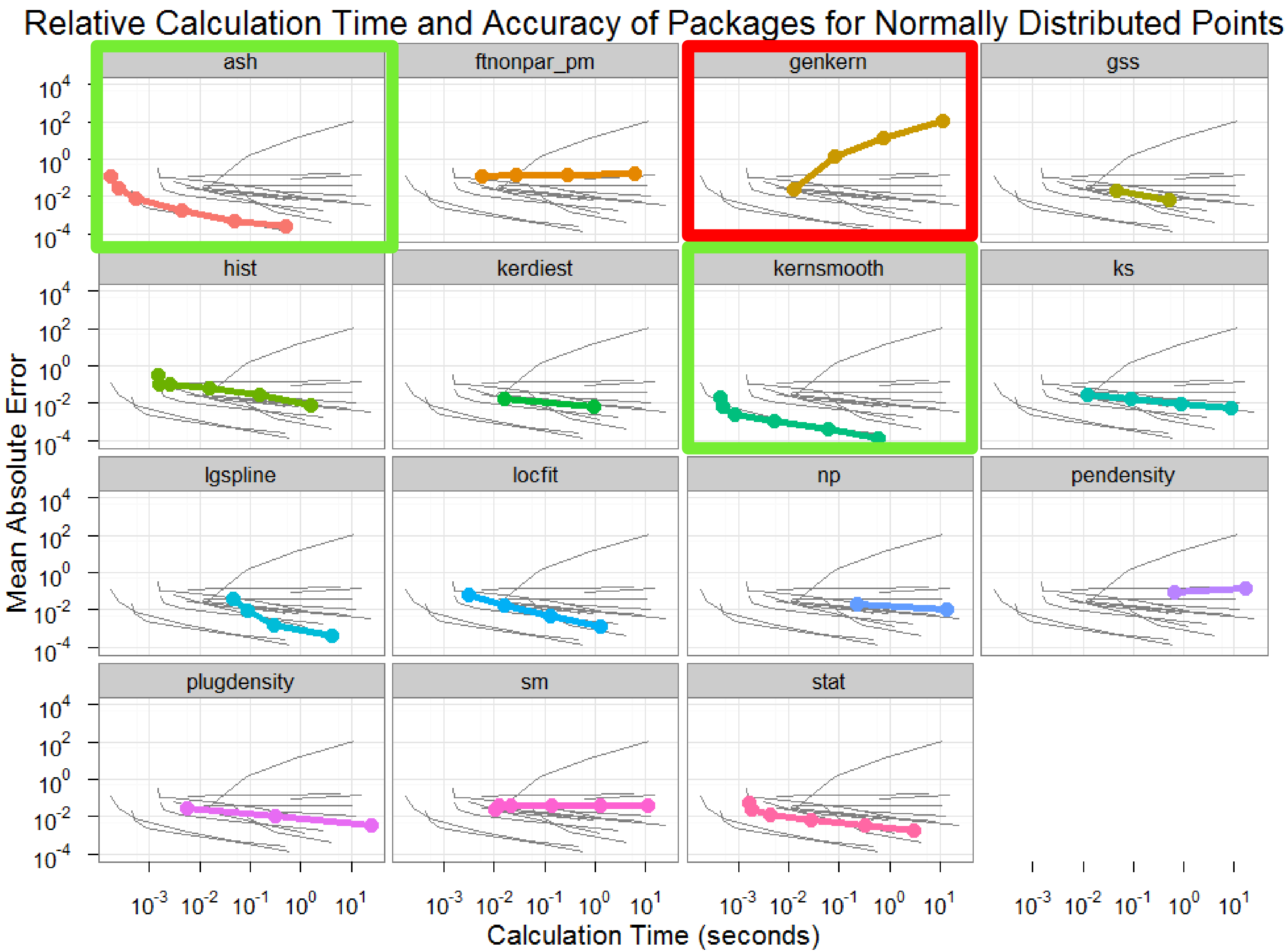




Density Estimation Absolute Error by Package
claw
norm
unif
Mean Absolute Error
Number of Data Points
100
1000
10000
1e+05
1e+06
1e+07
10^4
10^2
10^0
10^{-2}
10^{-4}
genkern
ftnonpar_pm
pendensity
lgspline
hist
stat
ash

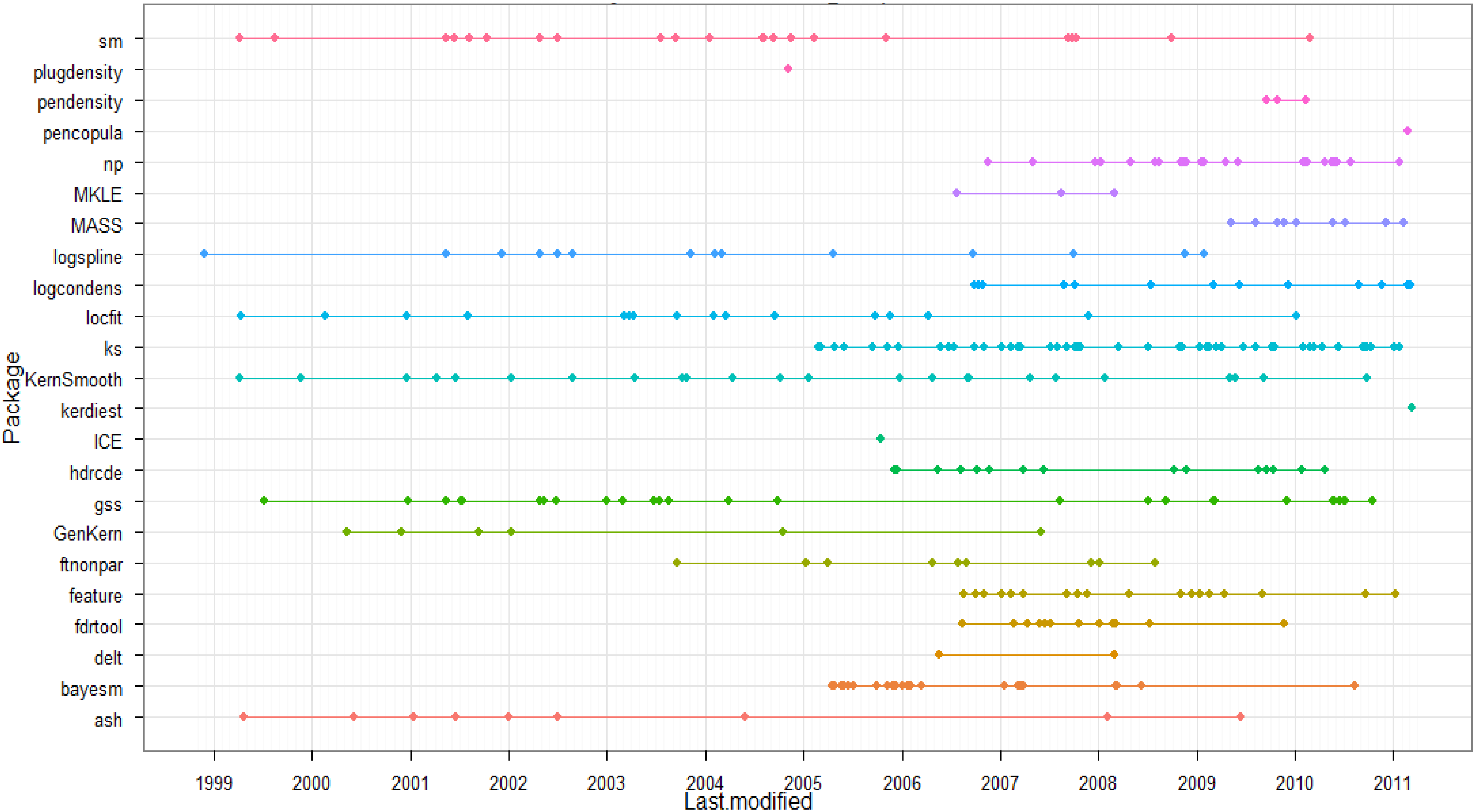
Additional Ideas

- Tradeoff between speed and accuracy
- Differences from uniform, normal and claw
- Impact of package update frequency
- Which theoretical approaches worked well?
  - Histograms, KDE's, other approaches



Relative Calculation Time and Accuracy of Packages for Normally Distributed Points
ash
ftnonpar_pm
genkern
gss
hist
kerdiest
kernsmooth
ks
lgspline
locfit
np
pendensity
plugdensity
sm
stat
Mean Absolute Error
Calculation Time (seconds)

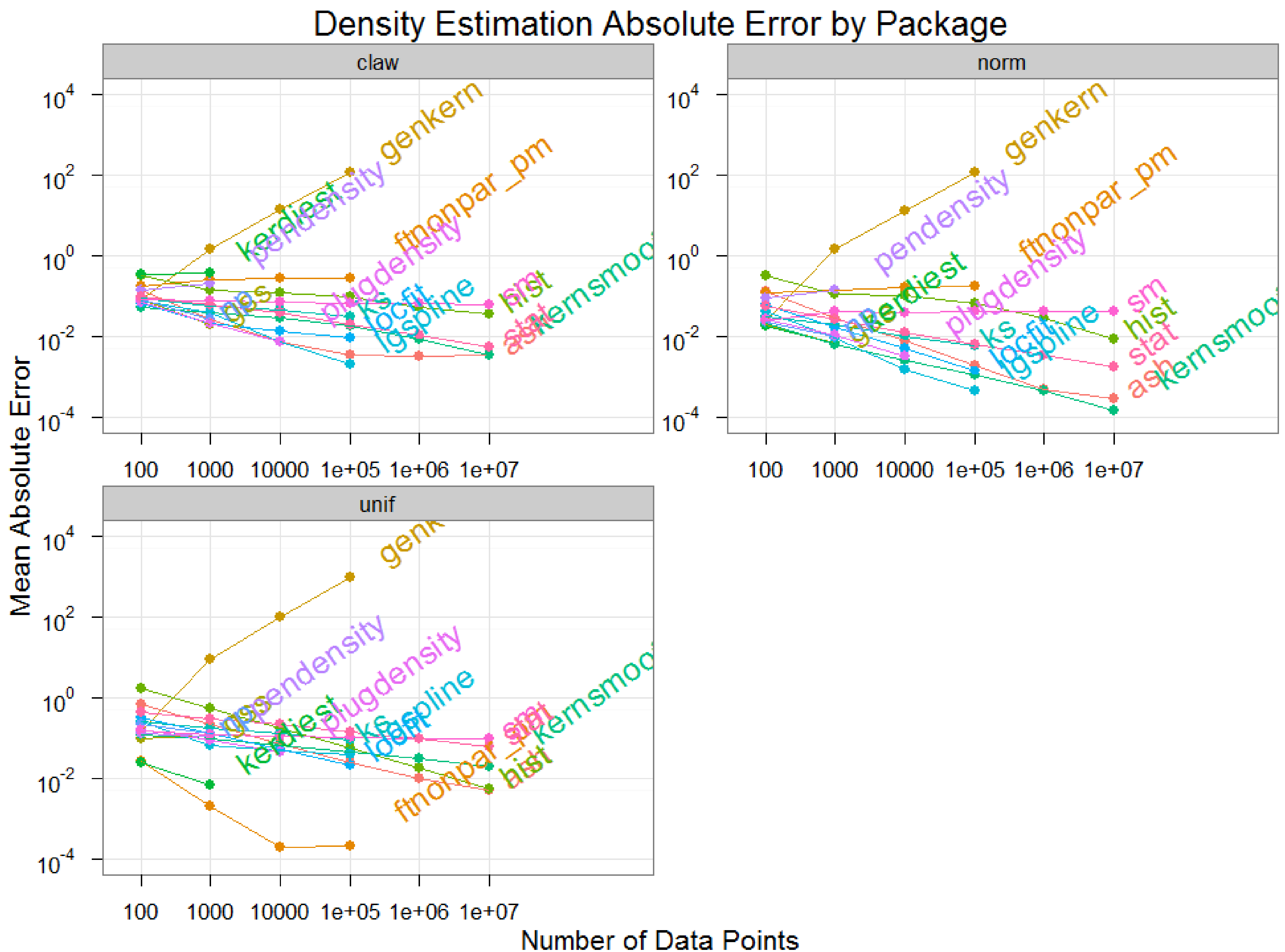
Package Updates

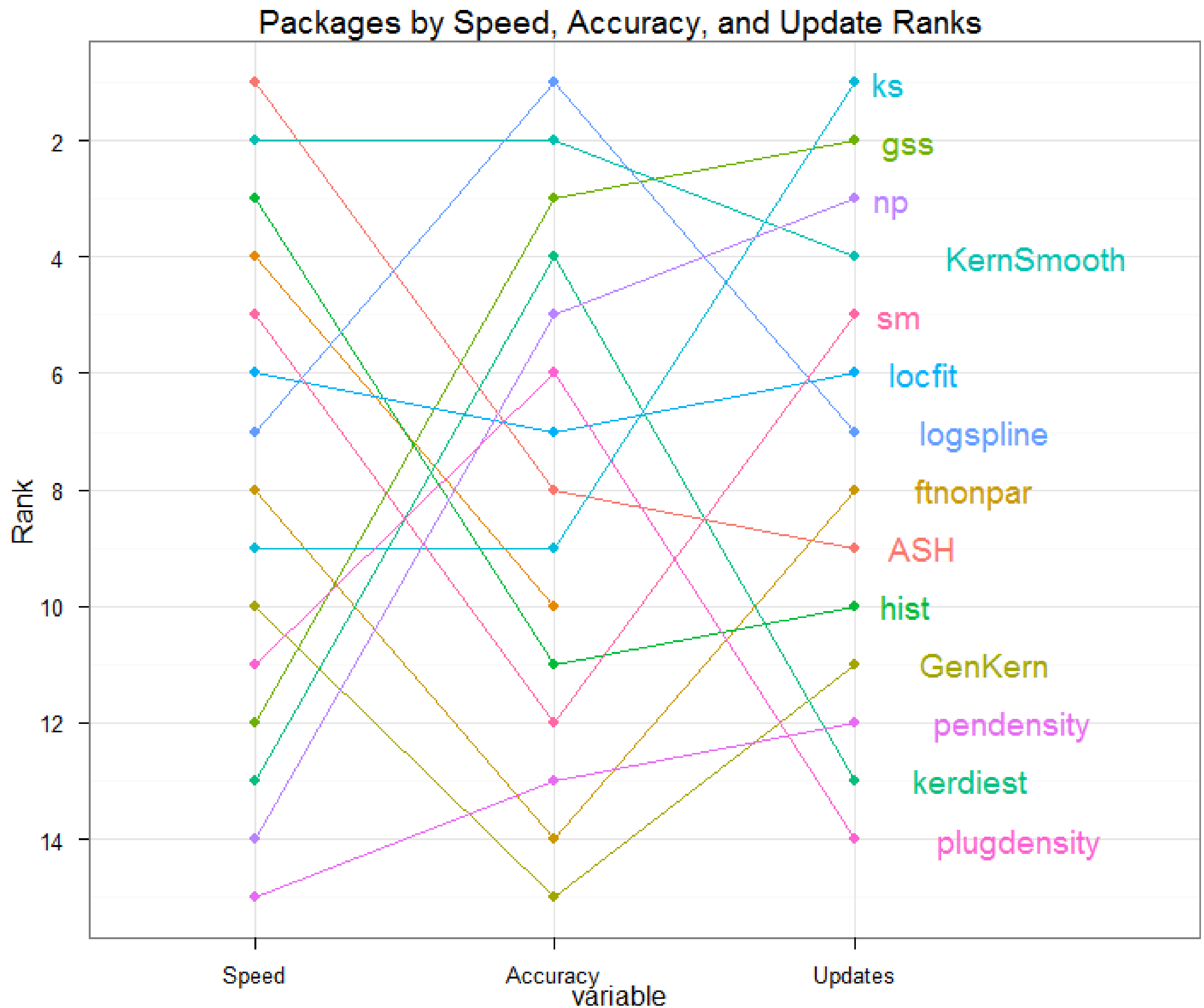


Conclusion

- Best packages are fast, accurate, and regularly updated without a speed/accuracy tradeoff
- Recommended packages: KernSmooth or ASH
  - KernSmooth uses binned KDE for speed
  - ASH uses averaged shifted histograms
- Extensions
  - Multivariate scenarios
  - Other kinds of density estimation
- Paper to be submitted to JSS

Optional Slides
(not used in presentation)


Density Estimation Absolute Error by Package
claw
norm
unif
Mean Absolute Error
Number of Data Points
10^4
10^2
10^0
10^-2
10^-4
100
1000
10000
1e+05
1e+06
1e+07
genkern
ftnonpar_pm
pendensity
kerdiest
plugdensity
gss
ks
locfit
lgspline
sm
hist
stat
ash
kernsmooth



Packages by Speed, Accuracy, and Update Ranks
Rank
2
4
6
8
10
12
14
Speed
Accuracy
Updates
variable
ks
gss
np
KernSmooth
sm
locfit
logspline
ftnonpar
ASH
hist
GenKern
pendensity
kerdiest
plugdensity

Package	Speed	Accuracy	Updates
ASH	1	8	9
density	4	10	
ftnonpar	8	14	8
GenKern	10	15	11
gss	12	3	2
hist	3	11	10
kerdiest	13	4	13
KernSmooth	2	2	4
ks	9	9	1
locfit	6	7	6
logspline	7	1	7
np	14	5	3
pendensity	15	13	12
plugdensity	11	6	14
sm	5	12	5

Time Vs. Accuracy Plot

```
name_scale <- scale_colour_hue("Package", limits = unique(all$name), legend = FALSE)
mpe_scale <- scale_y_log10("Mean Absolute Error", limits = c(10^-4, 10^4), breaks = 10^c(-4, -2, 0, 2, 4))

ggplot(norm, aes(med_time, med, colour = name)) +
    geom_line(data = transform(norm, id = name, name = NULL), colour = "grey50", aes(group = id)) +
    geom_line(size = 1.5) +
    geom_point(size = 3) +
    scale_x_log10("Calculation Time (seconds)") +
    scale_y_log10() +
    facet_wrap(~ name) +
    name_scale +
    mpe_scale +
    opts(title = "Relative Calculation Time and Accuracy of Packages for Normally Distributed Points")
```