Probabilistic forecasts for anomaly detection

Rob J Hyndman

3 July 2024

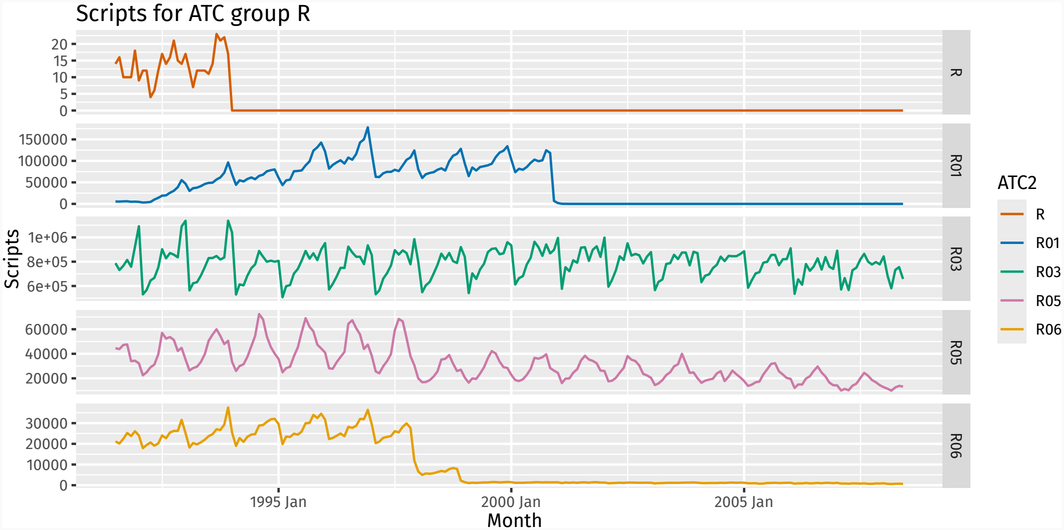
MONASH University

Australian PBS data

```
pbs
```

```
# A tsibble: 17,016 x 3 [1M]
# Key:       ATC2 [84]
   ATC2     Month Scripts
   <chr>    <mth>   <dbl>
 1 A01   1991 Jul   22615
 2 A01   1991 Aug   20443
 3 A01   1991 Sep   21389
 4 A01   1991 Oct   23746
 5 A01   1991 Nov   23477
 6 A01   1991 Dec   26316
 7 A01   1992 Jan   22041
 8 A01   1992 Feb   16393
 9 A01   1992 Mar   17207
10 A01   1992 Apr   18847
# i 17,006 more rows
```

Australian PBS data



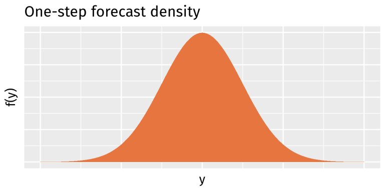
Main idea

- Estimate one-step forecast densities: $f(y_t | y_1, \dots, y_{t-1})$.
- Anomaly score: $s_t = -\log \hat{f}(y_t | y_1, \dots, y_{t-1})$.
- High anomaly scores indicate potential anomalies.
- Fit a Generalized Pareto Distribution to the top 5% of anomaly scores.
- Use the GPD to estimate the probability of each observation being an anomaly.

Anomaly score distribution

Suppose one-step forecasts are $N(\mu_t, \sigma^2)$.

So $f(y_t | y_1, \dots, y_{t-1}) = \phi\left(\frac{y_t - \mu_t}{\sigma}\right)$



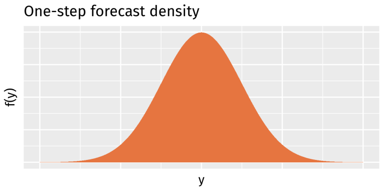
Anomaly score distribution

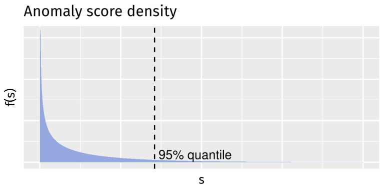
Suppose one-step forecasts are $N(\mu_t, \sigma^2)$.

So $f(y_t | y_1, \dots, y_{t-1}) = \phi\left(\frac{y_t - \mu_t}{\sigma}\right)$

Then $s_t = -\log \phi\left(\frac{y_t - \mu_t}{\sigma}\right) = \frac{1}{2}\left(\frac{y_t - \mu_t}{\sigma}\right)^2 + \frac{1}{2}\log(2\pi\sigma^2)$

So anomaly scores have distribution: $S \sim \frac{1}{2}\chi^2_1 + c$





Anomaly score distribution

Suppose one-step forecasts are $N(\mu_t, \sigma^2)$.

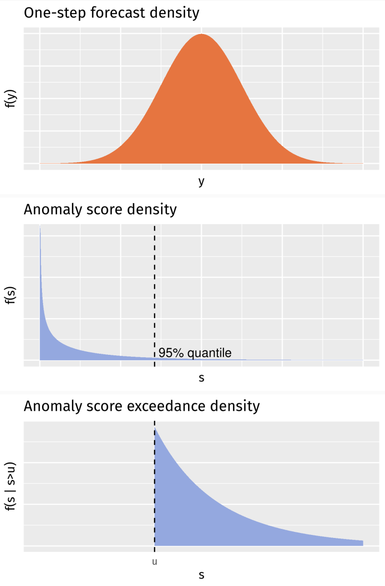
So $f(y_t | y_1, \dots, y_{t-1}) = \phi\left(\frac{y_t - \mu_t}{\sigma}\right)$

Then $s_t = -\log \phi\left(\frac{y_t - \mu_t}{\sigma}\right) = \frac{1}{2}\left(\frac{y_t - \mu_t}{\sigma}\right)^2 + \frac{1}{2}\log(2\pi\sigma^2)$

So anomaly scores have distribution: $S \sim \frac{1}{2}\chi^2_1 + c$

Conditional probability distribution of scores above threshold u is Generalized Pareto:

$$H(x) = P(S \le u + x \mid S > u) = 1 - (1 + \xi x / v)^{-1/\xi}$$

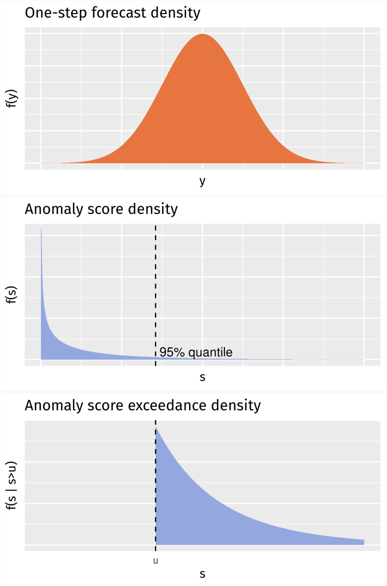


Anomaly score distribution

Extreme value theory shows that the Generalized Pareto distribution is a good approximation to the distribution of the largest anomaly scores for almost all possible forecast distributions.

Conditional probability distribution of scores above threshold u is Generalized Pareto:

$$H(x) = P(S \leq u + x \mid S > u) = 1 - (1 + \xi x/v)^{-1/\xi}$$



Fisher-Tippett-Gnedenko Theorem

$M_n = \max\{Y_1, \dots, Y_n\}$ where $Y_1, \dots, Y_n \sim$ iid F.

Under some conditions, as $n \to \infty$,
M_n converges in distribution to

Weibull: when F has a finite upper bound (e.g., Uniform)

Gumbel: when F has exponential tails (e.g., Normal or Gamma)

Fréchet: when F has heavy tails (e.g., Pareto or Weibull)

Generalized Pareto distribution

Peaks Over Threshold (POT): extremes are observations $> u$.

Probability distribution of extremes::

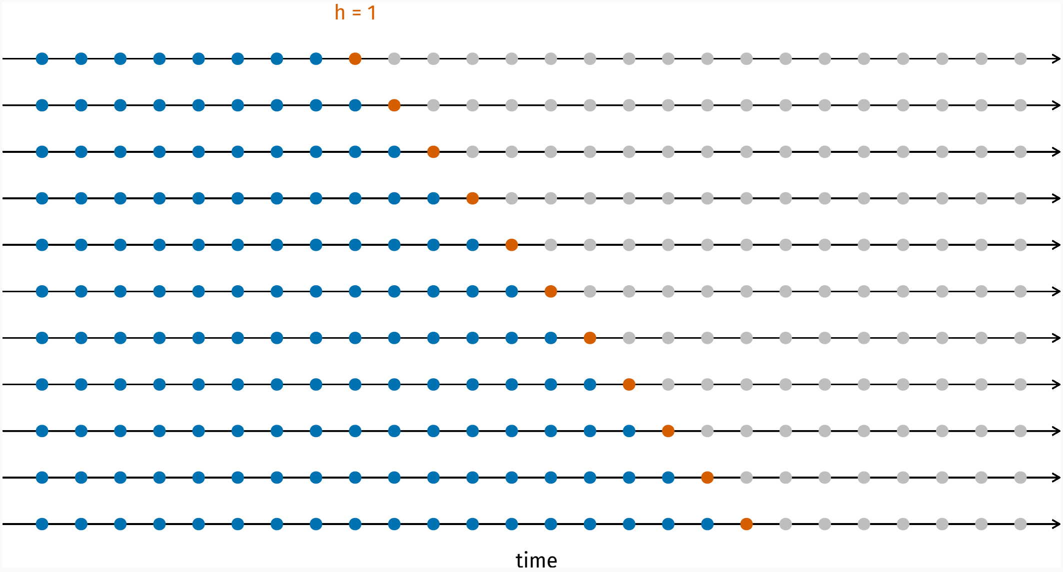
$$H(y) = P\{Y \leq u + y \mid Y > u\} = \frac{F(u + y) - F(u)}{1 - F(u)}.$$

When F satisfies conditions of FTG theorem, then H is **Generalized Pareto Distribution** (GPD):

$$H(y) \approx 1 - (1 + \xi y/v)^{-1/\xi},$$

where $\{y : y > 0 \text{ and } (1 + \xi y)/v > 0\}$.

Rolling origin forecasts



Rolling origin forecasts

```r
pbs_stretch <- stretch_tsibble(pbs, .step = 1, .init = 36)
```

```
# A tsibble: 1,684,884 x 4 [1M]
# Key:       .id, ATC2 [14,076]
   ATC2     Month Scripts   .id
   <chr>    <mth>   <dbl> <int>
 1 A01   1991 Jul   22615     1
 2 A01   1991 Aug   20443     1
 3 A01   1991 Sep   21389     1
 4 A01   1991 Oct   23746     1
 5 A01   1991 Nov   23477     1
 6 A01   1991 Dec   26316     1
 7 A01   1992 Jan   22041     1
 8 A01   1992 Feb   16393     1
 9 A01   1992 Mar   17207     1
10 A01   1992 Apr   18847     1
# i 1,684,874 more rows
```

Rolling origin forecasts

```
pbs_fit <- pbs_stretch |> model(ets = ETS(Scripts))
```

```
# A mable: 14,076 x 3
# Key:     .id, ATC2 [14,076]
     .id ATC2            ets
   <int> <chr>       <model>
 1     1 A01    <ETS(M,N,A)>
 2     1 A02    <ETS(M,A,M)>
 3     1 A03    <ETS(M,A,M)>
 4     1 A04    <ETS(M,N,M)>
 5     1 A05   <ETS(A,Ad,N)>
 6     1 A06    <ETS(M,N,M)>
 7     1 A07    <ETS(M,A,M)>
 8     1 A09    <ETS(M,A,M)>
 9     1 A10    <ETS(M,A,M)>
10     1 A11    <ETS(M,A,M)>
# i 14,066 more rows
```

Rolling origin forecasts

```r
pbs_fc <- forecast(pbs_fit, h = 1)
```

```
# A fable: 14,076 x 4 [1M]
# Key:     .id, ATC2 [14,076]
     .id ATC2     Month               Scripts
   <int> <chr>    <mth>                <dist>
 1     1 A01   1994 Jul  N(22722, 2441206)
 2     1 A02   1994 Jul N(588422, 1.1e+09)
 3     1 A03   1994 Jul  N(84529, 1.9e+07)
 4     1 A04   1994 Jul  N(70220, 1.5e+07)
 5     1 A05   2003 Jul     N(1372, 13768)
 6     1 A06   1994 Jul  N(30624, 5537439)
 7     1 A07   1994 Jul  N(78305, 1.5e+07)
 8     1 A09   1994 Jul     N(3658, 28241)
 9     1 A10   1994 Jul N(166969, 5.4e+07)
10     1 A11   1994 Jul  N(30575, 2947627)
# i 14,066 more rows
```

Finding anomalies

```
pbs_scores <- pbs_fc |>
  left_join(pbs |> rename(actual = Scripts), by = c("ATC2", "Month")) |>
  mutate(
    s = -log_likelihood(Scripts, actual), # Density scores
    prob = lookout(density_scores = s)    # Probability not an anomaly
  )
```

```
# A fable: 14,076 x 7 [1M]
# Key:     .id, ATC2 [14,076]
     .id ATC2      Month             Scripts actual     s  prob
   <int> <chr>     <mth>              <dist>  <dbl> <dbl> <dbl>
 1     1 A01    1994 Jul  N(22722, 2441206)  20854  8.99 1
 2     1 A02    1994 Jul N(588422, 1.1e+09) 516122 13.8  0.583
 3     1 A03    1994 Jul  N(84529, 1.9e+07)  80471  9.73 1
 4     1 A04    1994 Jul  N(70220, 1.5e+07)  66125  9.74 1
 5     1 A05    2003 Jul     N(1372, 13768)   1468  6.02 1
 6     1 A06    1994 Jul  N(30624, 5537439)  29194  8.87 1
 7     1 A07    1994 Jul  N(78305, 1.5e+07)  68542 12.4  1
 8     1 A09    1994 Jul     N(3658, 28241)   3320  8.07 1
```

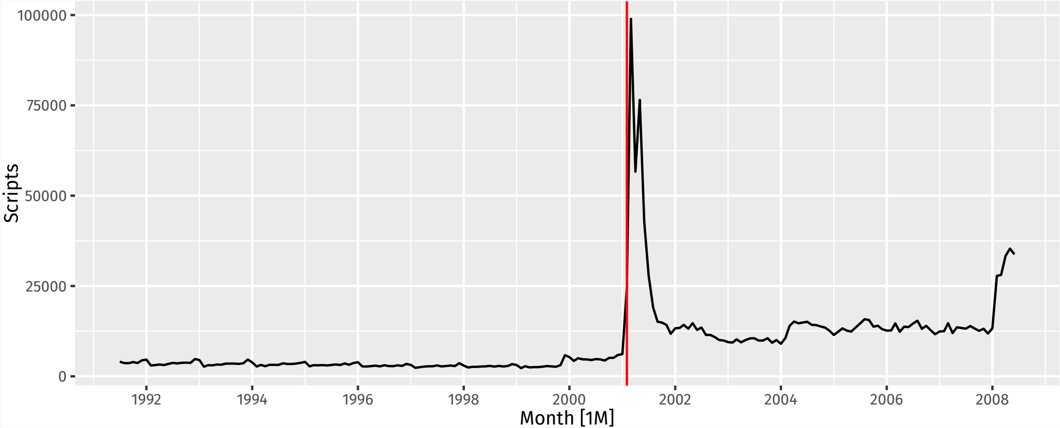
Finding anomalies

```
pbs_scores |> arrange(prob)
```

```
# A tsibble: 14,076 x 7 [1M]
# Key:       .id, ATC2 [14,076]
     .id ATC2     Month           Scripts actual     s     prob
   <int> <chr>    <mth>            <dist>  <dbl> <dbl>    <dbl>
 1    80 N07   2001 Feb  N(4261, 136066)  24616 1529. 0.000925
 2    44 R06   1998 Feb  N(-1222, 17510)   4986 1106. 0.00128
 3   146 P01   2006 Aug       N(26, 5.6)    129  951. 0.00148
 4    81 N07   2001 Mar N(8484, 4549377)  98942  908. 0.00156
 5    18 L03   1996 Dec      N(329, 536)   1231  763. 0.00185
 6   131 D11   2005 May      N(136, 178)    596  598. 0.00236
 7    24 C05   1996 Jun        N(5, 1.8)     50  567. 0.00249
 8   141 P01   2006 Mar     N(506, 1617)   1505  313. 0.00455
 9    55 R06   1999 Jan   N(-835, 12055)   1452  223. 0.00648
10    57 D05   1999 Mar    N(783, 11837)   2789  176. 0.00832
# i 14,066 more rows
```

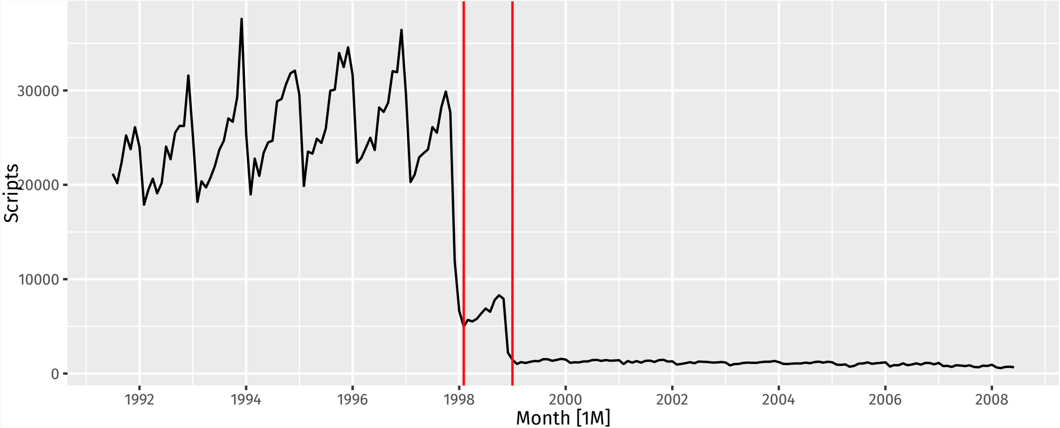
Finding anomalies

```
pbs |> filter(ATC2 == "N07") |> autoplot() +
  scale_x_yearmonth(date_breaks = "2 years", date_labels = "%Y") +
  geom_vline(xintercept = as.Date("2001-02-01"), color = "red")
```



Finding anomalies

```
pbs ▷ filter(ATC2 == "R06") ▷ autoplot() +
  scale_x_yearmonth(date_breaks = "2 years", date_labels = "%Y") +
  geom_vline(xintercept = as.Date(c("1998-02-01","1999-01-01")), color = "red")
```



Online anomaly detection

- Demonstrate using weird package with (a) univariate models for tourism data; and (b) univariate models for age-specific time series from French mortality.