Report

Anthony Ebert

22/08/2019

We are interested in finding the closest mechanistic model $q(\cdot|\lambda)$, parameterised by λ to a known statistical model $p(\cdot|\theta)$, where θ is known.

$$\begin{aligned} KL[q(\cdot)||p(\cdot|\theta)] &= \sum_{y\in\mathcal{Y}} q(y)\log\frac{q(y)}{p(y|\theta)} \\ &= \log[z(\theta)] - \sum_{y\in\mathcal{Y}} q(y)\left[\sum_i \theta_i s_i(y)\right] + \sum_{y\in\mathcal{Y}} q(y)\log q(y) \end{aligned}$$

Entropy estimation

I use the following non-parameteric estimator of entropy (Vu, Yu, and Kass 2007):

$$\begin{aligned} \tilde{H} &:= -\sum_k \frac{\tilde{p}_k \log \tilde{p}_k}{1-(1-\tilde{p}_k)^n} \\ \tilde{p}_k &:= \hat{C}\hat{p}_k \\ \hat{C} &:= 1 - \frac{\#\{k|n_k = 1\}}{\sum_k n_k} \\ \hat{p}_k &:= \frac{n_k}{n} \end{aligned}$$

For instance

```
set.seed(1)

library(StartNetwork)

x <- rbinom(1000, 10, 0.1)
head(x)
```

```
## [1] 0 1 1 2 0 2
```

```
mean(-log(dbinom(x, size = 10, 0.1)))
```

```
## [1] 1.292132
```

```
entropy_calc(x)
```

```
## [1] 1.295627
```

```
x <- rbinom(1000, 10, 0.2)
head(x)
```

```
## [1] 2 3 2 4 1 0
```

```
mean(-log(dbinom(x, size = 10, 0.2)))
```

```
## [1] 1.63193
```

```
entropy_calc(x)
```

```
## [1] 1.625031
```

The relative importance of the likelihood and entropy

```
library(parallel)
cl <- makeCluster(detectCores())

param_range_large <- rep(seq(0.025, 0.975, by = 0.025), 100)

x1 <- parSapply(cl, param_range_large, er_KL, pl = 0.3, include_entropy = TRUE, replicates = 1000)
x2 <- parSapply(cl, param_range_large, er_KL, pl = 0.3, include_entropy = FALSE, replicates = 1000)

param_range_small <- rep(seq(0.275, 0.325, by = 0.005), 100)

x3 <- parSapply(cl, param_range_small, er_KL, pl = 0.3, include_entropy = TRUE, replicates = 1000)
x4 <- parSapply(cl, param_range_small, er_KL, pl = 0.3, include_entropy = FALSE, replicates = 1000)
```

```
library(ggplot2)


df <- data.frame(x = x1, parameter = param_range_large)

ggplot(df) + aes(x = parameter, y = x, group = parameter) + geom_boxplot()
```

```
df <- data.frame(x = x2, parameter = param_range_large)

ggplot(df) + aes(x = parameter, y = x, group = parameter) + geom_boxplot()
```

```
df <- data.frame(x = x3, parameter = param_range_small)

ggplot(df) + aes(x = parameter, y = x, group = parameter) + geom_boxplot()
```

```
df <- data.frame(x = x4, parameter = param_range_small)

ggplot(df) + aes(x = parameter, y = x, group = parameter) + geom_boxplot()
```

Attempt with preferential attachment model

The preferential attachment model has the following likelihood function:

$$P(k|\rho) = \frac{(\rho - 1)\Gamma(k)\Gamma(\rho)}{\Gamma(k + \rho)}$$

In this case there is no simple set of summary statistics.

```
x <- sapply(seq(2.5, 5.5, by = 0.1), ba_KL, pl = 4.5)

ggplot2::qplot(seq(2.5, 5.5, by = 0.1),x, xlab = "power parameter", ylab = "KL divergence") + ggplot2::
```

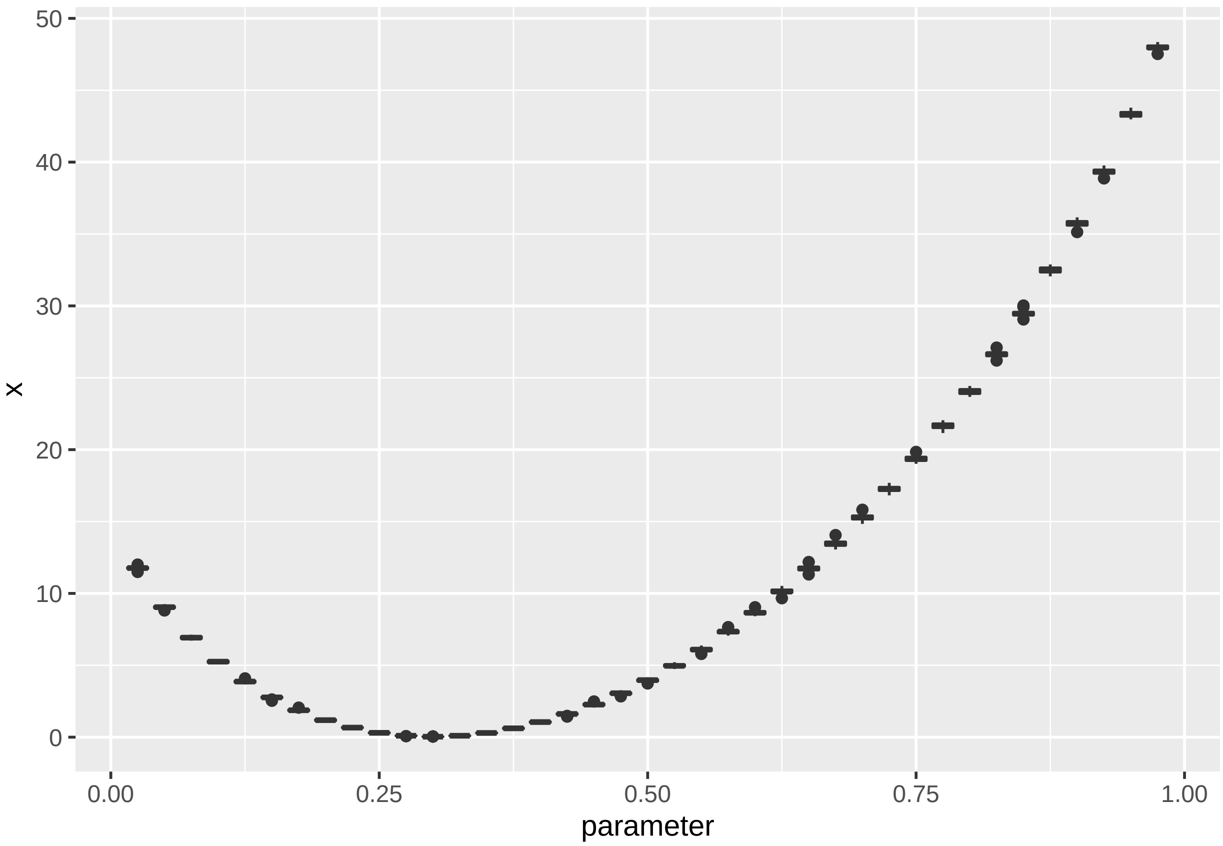


Figure 1: KL divergence calculation with entropy

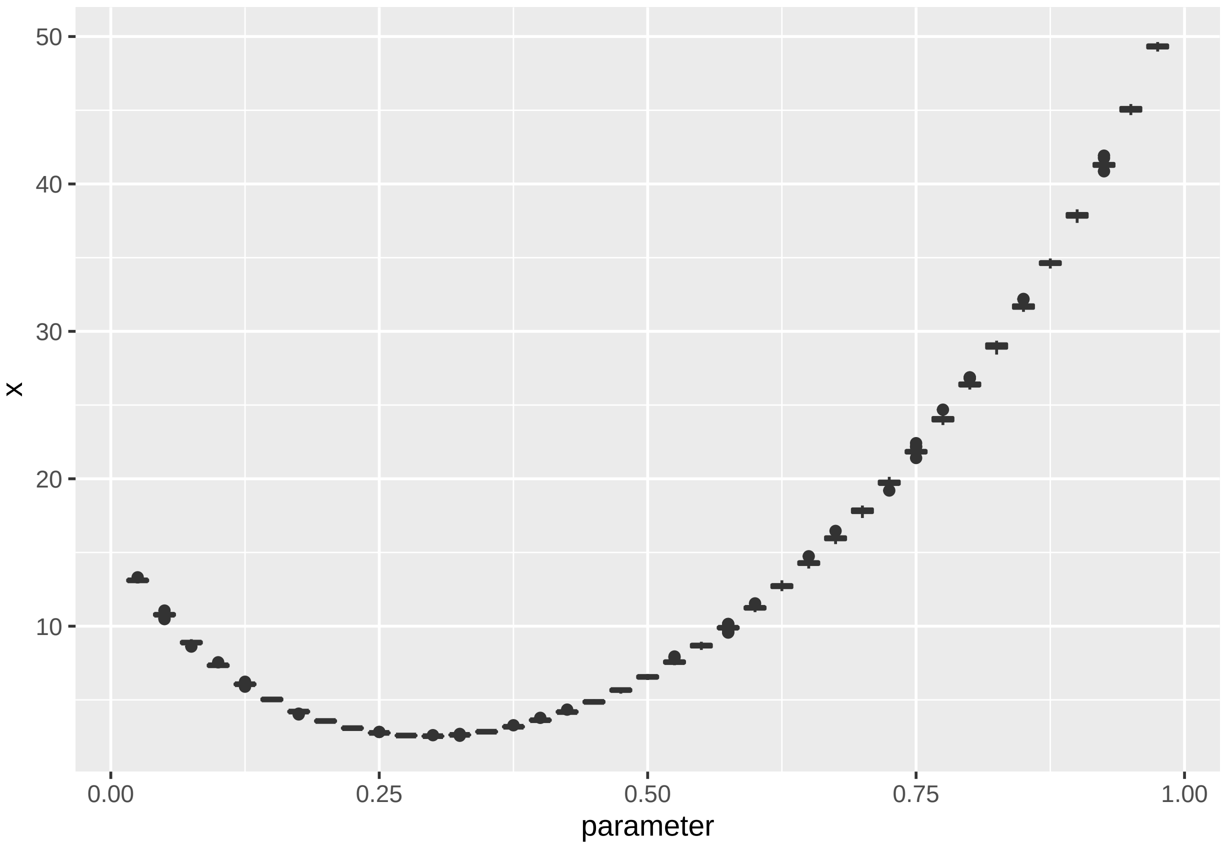


Figure 2: KL divergence calculation without entropy

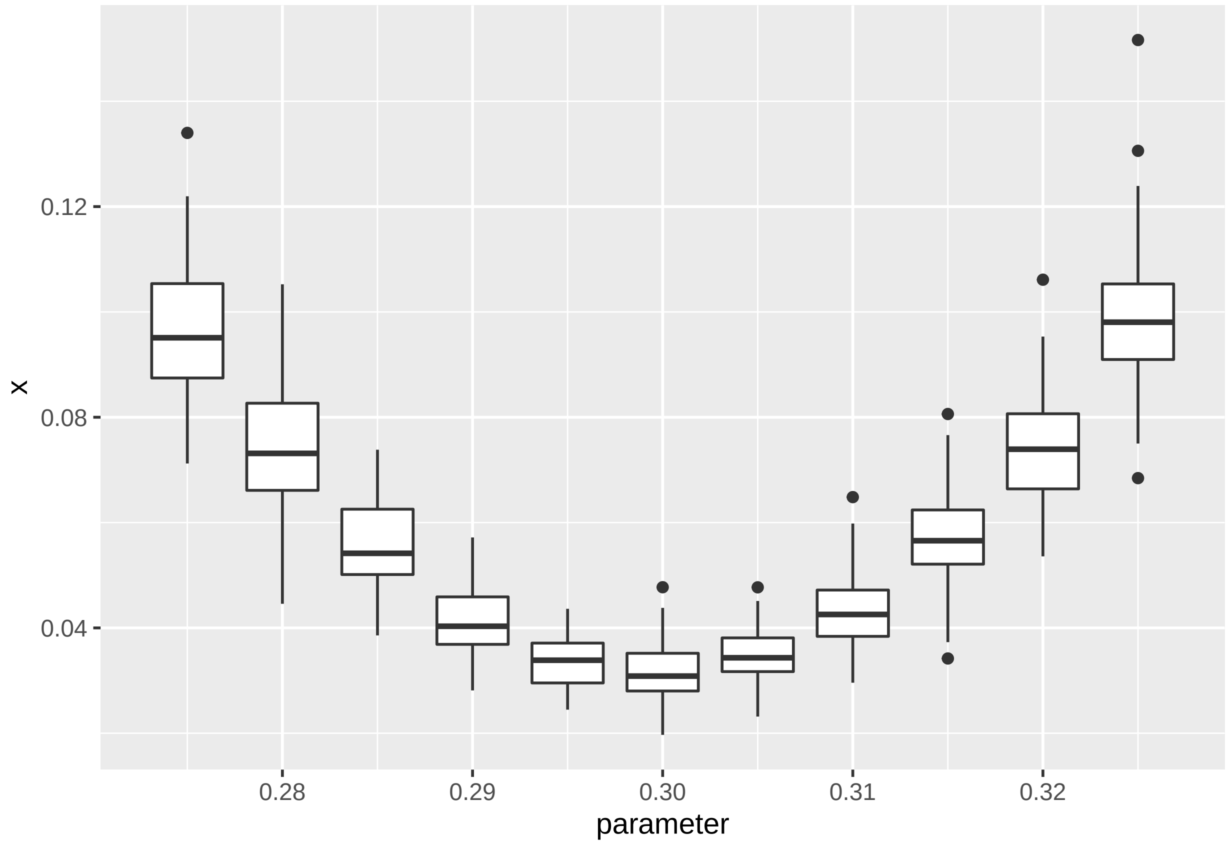


Figure 3: KL divergence calculation with entropy

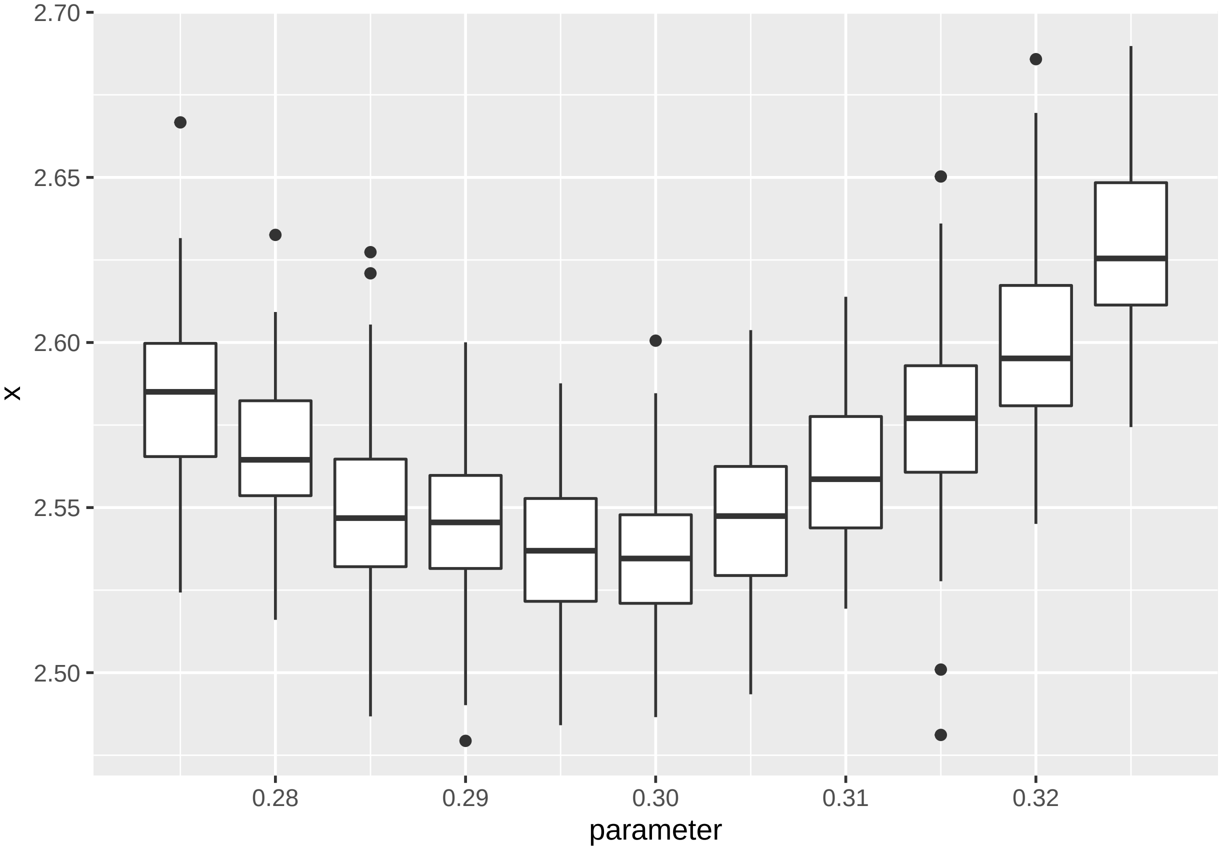
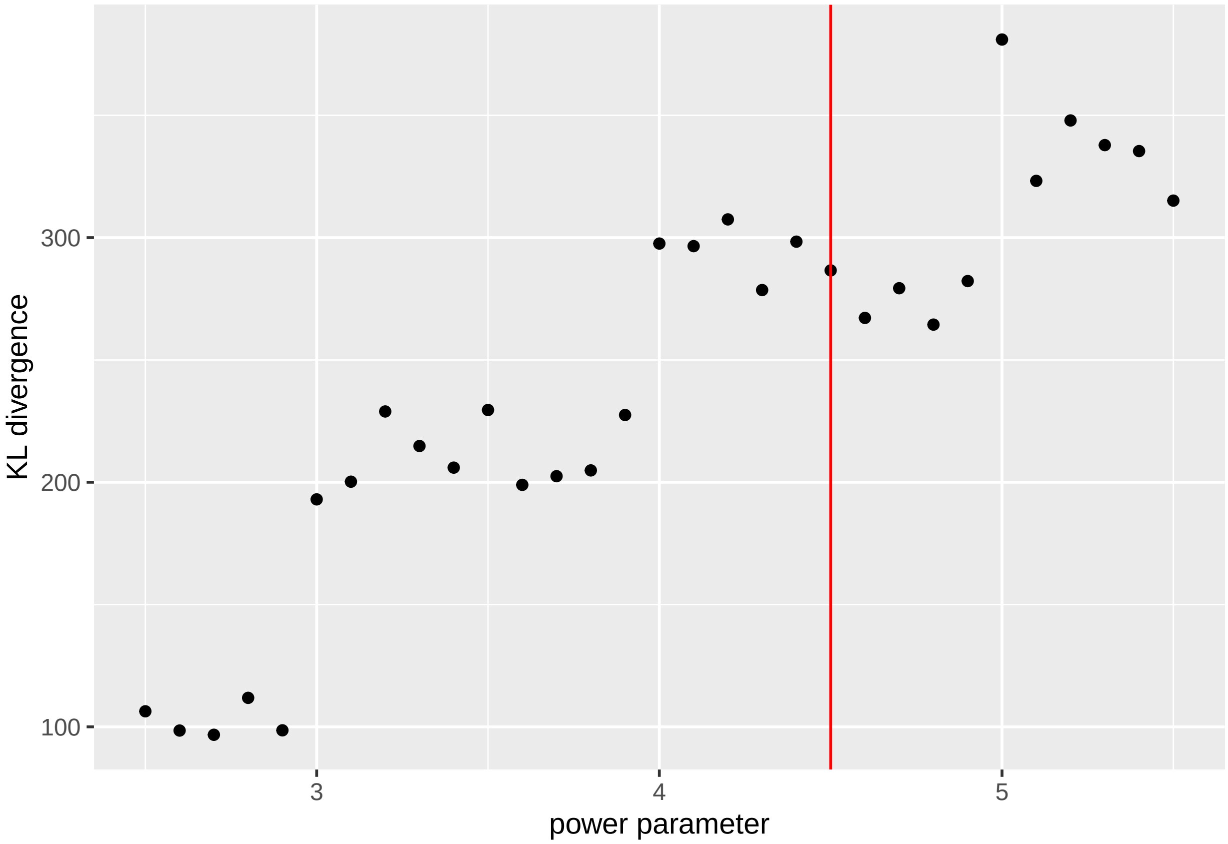


Figure 4: KL divergence calculation without entropy



Vu, Vincent Q, Bin Yu, and Robert E Kass. 2007. "Coverage-Adjusted Entropy Estimation." *Statistics in Medicine* 26 (21). Wiley Online Library: 4039–60.