STAT-S676

Assignment 1

John Koo

See source code here: https://github.com/johneverettkoo/stats-hw

Problem 1

AIC is given by $-2\log(L(\hat{\theta}|y)) + 2K$ where L is the likelihood function. We are given that $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$. Then $\varepsilon = Y - X\beta \sim (0, \sigma^2 I)$. Therefore, the pdf of ε is a multivariate normal $f(\varepsilon) = \big(\frac{1}{2\pi\sigma^2}\big)^{n/2} e^{-\frac{1}{2\sigma^2}\varepsilon^T\varepsilon}$. $\varepsilon^T\varepsilon = RSS = n\hat{\sigma}^2 \approx n\sigma^2$. (We sub in the estimate for the parameter.) Then we get $L = \big(\frac{1}{2\pi\sigma^2}\big)^{n/2} e^{-\frac{n}{2}}$. Plugging this in for L, we get:

$$
\begin{aligned}
AIC &= -2\log\left(\Big(\frac{1}{2\pi\sigma^2}\Big)^{n/2} e^{-\frac{n}{2}}\right) + 2K \\
&= -2\bigg(-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{n}{2}\bigg) + 2K \\
&= n\log(2\pi) + n\log(\sigma^2) + n + 2K
\end{aligned}
$$

We can ignore the two terms $n\log(2\pi)$ and n since they are fixed for some particular dataset.

Problem 2

We have the following components for y, β, and σ^2:

$$f(y|\beta, \sigma^2) = \Big(\frac{1}{2\pi\sigma^2}\Big)^{n/2} e^{-\frac{(y-X\beta)^T(y-X\beta)}{2\sigma^2}}$$

$$f(\beta|\sigma^2) = \frac{\left|X^T X\right|^{1/2}}{(2\pi n\sigma^2)^{p/2}} e^{-\frac{(X\beta)^T(X\beta)}{2n\sigma^2}}$$

$$f_{\sigma^2}(\sigma^2) = \Big(\frac{1}{2\pi(\sigma^2)^3}\Big)^{1/2} e^{-\frac{1}{2\sigma^2}}$$

Then

$$f(y, \beta, \sigma^2) = (y|\beta, \sigma^2) \times f(\beta|\sigma^2) \times f_{\sigma^2}(\sigma^2)$$

And to compute the marginal for Y:

$$f_Y(y) = \int_{\beta, \sigma^2} f(y, \beta, \sigma^2) d\beta d\sigma^2$$

The strategy is to integrate out y and β by completing the square. Combining the exponents, we obtain:

$$\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\left(\frac{\left|X^TX\right|}{(2\pi n\sigma^2)^p}\right)^{1/2}\left(\frac{1}{2\pi(\sigma^2)^3}\right)^{1/2}\exp\left(-\frac{1}{2\sigma^2}\Big((y-X\beta)^T(y-X\beta)+\frac{1}{n}\beta^TX^TX\beta+1\Big)\right)$$

Then rearranging the terms inside the exponent:

$$-\frac{1}{2\sigma^2}\left((\beta-\hat{\beta}\frac{n}{n+1})^T\left(X^TX(1+\frac{1}{n})\right)(\beta-\hat{\beta}\frac{n}{n+1})+y^T\left(I-\frac{n}{n+1}H\right)y+1\right)$$

Where $\hat{\beta}=(X^TX)^{-1}X^Ty$ and $H=X(X^TX)^{-1}X^T$.

Then we can see that $\Sigma_\beta = \sigma^2\frac{n}{n+1}(X^TX)^{-1}$. Then integrating w.r.t. β, we get a factor of $\left((2\pi)^p(\sigma^2)^p(\frac{n}{n+1})^p/\left|X^TX\right|\right)^{1/2}$, which leaves us with:

$$f_{Y,\sigma^2}=\left(\frac{1}{2\pi\sigma^2}\right)^{n/2+1/2}\left(\frac{1}{n+1}\right)^{p/2}(\sigma^2)^{-\frac{3}{2}-\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}\Big(y^T\left(I-\frac{n}{n+1}H\right)y+1\Big)\right)$$

Integrating out σ^2, we obtain:

$$\left(\frac{1}{2\pi}\right)^{\frac{n+1}{2}}\left(\frac{1}{n+1}\right)^{p/2}\frac{\Gamma(\frac{n+1}{2})}{\left(\frac{1}{2}y^T(I-\frac{n}{n+1}H)y+1/2\right)^{\frac{n+1}{2}}}$$

Problem 3

TIC is defined by $-2\log L+2tr(J(\theta)I(\theta^{-1}))$. $L(\theta|y)=g(y|\theta)$ where $I=\nabla_\theta\nabla_\theta^Tl$ and $J=(\nabla_\theta)^T(\nabla_\theta)$. In terms of partial derivatives, this becomes:

$$I=-E_{truth}\begin{bmatrix}\frac{\partial^2 l}{\partial\beta\partial\beta^T} & \frac{\partial^2 l}{\partial\beta\partial\sigma^2}\\ \frac{\partial^2 l}{\partial\sigma^2\partial\beta^T} & \frac{\partial^2 l}{\partial(\sigma^2)^2}\end{bmatrix}$$

$$J=E_{truth}\begin{bmatrix}(\partial_\beta l)(\partial_\beta l)^T & (\partial_\beta l)(\partial_{\sigma^2} l)^T\\ (\partial_{\sigma^2} l)(\partial_{\beta^2} l)^T & (\partial_{\sigma^2} l)(\partial_{\sigma^2} l)^T\end{bmatrix}$$

Then computing the partial derivatives:

$$\frac{\partial l}{\partial\beta}=\frac{(y-X\beta)^TX}{\sigma^2}$$

$$\frac{\partial l}{\partial\sigma^2}=-\frac{1}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}(y-X\beta)^T(y-X\beta)$$

$$\frac{\partial^2 l}{\partial\beta\beta^T}=-\frac{X^TX}{\sigma^2}$$

$$\frac{\partial^2 l}{\partial\beta\sigma^2}=-\frac{(y-X\beta)^TX}{(\sigma^2)^2}$$

$$\frac{\partial^2 l}{\partial(\sigma^2)^2}=\frac{1}{2(\sigma^2)^2}-\frac{1}{(\sigma^2)^3}(y-X\beta)^T(y-X\beta)$$

Noting that $y\sim\mathcal{N}(\mu,\sigma^2I)$, taking the expectations, we arrive at:

$$I = \frac{1}{\sigma^2}\begin{bmatrix} X^TX & \frac{X^T(\mu-X\beta)}{\sigma^2} \\ \frac{(\mu-X\beta)^TX}{\sigma^2} & -\frac{1}{2\sigma^2}+\frac{\sigma^2+(\mu-X\beta)^T(\mu-X\beta)}{(\sigma^2)^2} \end{bmatrix}$$

Where the numerator of the second term of $[I]_{22}$ was derived as follows: $(y-X\beta)^T(y-X\beta) = (y-\mu+\mu-X\beta)^T(y-\mu+\mu-X\beta) = (y-\mu)^T(y-\mu)+(\mu-X\beta)^T(\mu-X\beta)+2(y-\mu)^T(\mu-X\beta)$. Then taking the expectation of this, the first term becomes σ^2, the second term stays the same, and the last term goes to 0 since $E[y]=\mu$.

In order to compute J, we need the following:

$$[J]_{11} = E\Big[\frac{1}{(\sigma^2)^2}X^T(y-X\beta)(y-X\beta)^TX\Big]$$

The employing the same trick as before, the middle part becomes:

$$(y-\mu+\mu-X\beta)(y-\mu+\mu-X\beta)^T$$
$$= (y-\mu)(y-\mu)^T+(y-\mu)(\mu-X\beta)^T+(\mu-X\beta)(y-\mu)^T+(\mu-X\beta)(\mu-X\beta)^T$$

Under the expectation, the middle two terms go to 0 since $E[y]=\mu$. The first term also turns into $\sigma^2 I$. Then we get:

$$[J]_{11} = \frac{1}{(\sigma^2)^2}X^T(\sigma^2 I+(\mu-X\beta)(\mu-X\beta)^T)X$$

For $[J]_{22}$, we compute $\left(\frac{\partial^2 l}{\partial(\sigma^2)^2}\right)^2$ then take the expected value. We can say that the odd powers go to zero under the expectation since the normal is symmetric. Then we get:

$$\frac{1}{4(\sigma^2)^2}\Bigg(1-\frac{2}{\sigma^2}\big(\sigma^2+(\mu-X\beta)^2\big)+\frac{1}{(\sigma^2)^2}\big(3(\sigma^2)^2+6\sigma^2(\mu-X\beta)^T(\mu-X\beta)+((\mu-X\beta)^T(\mu-X\beta)^T)^2\big)\Bigg)$$

For the $[J]_{12} = (\frac{\partial l}{\partial\beta})(\frac{\partial l}{\partial\sigma^2})^T$ term:

$$(\frac{\partial l}{\partial\beta})(\frac{\partial l}{\partial\sigma^2})^T = \frac{1}{2(\sigma^2)^2}(y-X\beta)^TX\Bigg(-1+\frac{1}{\sigma^2}(y-X\beta)^T(y-X\beta)\Bigg)$$

Taking the expectation of this *should* yield:

$$\frac{1}{2(\sigma^2)^2}\Bigg((\mu-X\beta)+\frac{1}{\sigma^2}(3\sigma^2(\mu-X\beta)+(\mu-X\beta)^T(\mu-X\beta)(\mu-X\beta)^T\Bigg)$$

$[J]_{21}$ is just the transpose of $[J]_{12}$.

Problem 4

For the sake of CPU time, I am going to omit intercept-less models.

```r
t0 <- Sys.time()

# --- setup. --- #

# packages, etc.
import::from(magrittr, `%>%`, `%<>%`)
library(ggplot2)
theme_set(theme_bw())
dp <- loadNamespace('dplyr')
import::from(parallel, mclapply, detectCores)
import::from(purrr, flatten)
import::from(viridis, scale_color_viridis)

# get the data
load('~/dev/stats-hw/stat-s676/diabetes3.Rdata')

# precompute stuff
y <- diabetes3$y
x <- diabetes3 %>%
  dp$select(-y) %>%
  as.matrix()
xt.x <- crossprod(x, x)
xt.y <- crossprod(x, y)
yt.y <- crossprod(y)

# parameters
tol <- 1e-12
alpha. <- 1  # to avoid conflict with the alpha function in R
beta. <- 1  # to avoid conflict with the beta function in R
n <- length(y)
p <- ncol(x)
n.mod <- 2 ** p  # number of models
mc.offset <- 1  # number of cores to not use
p.names <- colnames(x)
sample.frac <- .01  # number of rows to use for visualization

# precompute some more stuff
log.samp.const <- -n / 2 * log(2 * pi * exp(1) / n)
log.marg.const <- lgamma((n + alpha.) / 2) - n * log(pi) - lgamma(alpha. / 2)

# list of models
mod.list <- lapply(seq(p), function(i) {
  combn(p.names, i, simplify = FALSE)
}) %>%
  flatten()

# set up parallel stuff
options(mc.cores = detectCores() - mc.offset)

# group model indices for parallelization
mod.chunks <- split(seq_along(mod.list), seq(detectCores() - mc.offset))

# parallel computation
```

```r
out.df <- mclapply(mod.chunks, function(chunk) {
  lapply(chunk, function(i) {
    # which model?
    mod.loc <- mod.list[[i]]
    p.loc <- length(mod.loc)

    xt.x.loc <- xt.x[mod.loc, mod.loc]
    xt.y.loc <- xt.y[mod.loc, ]
    eig.loc <- eigen(xt.x.loc, symmetric = TRUE)
    logical.loc <- ((eig.loc$values / eig.loc$values[1]) > tol)
    inv.eig.vals.loc <- ifelse(logical.loc, 1 / eig.loc$values, rep(0, p.loc))
    rank.loc <- sum(logical.loc)
    xt.x.loc.inv <-
      eig.loc$vectors %*%
      diag(inv.eig.vals.loc, p.loc, p.loc) %*%
      t(eig.loc$vectors)
    hat.beta.loc = crossprod(xt.x.loc.inv, xt.y.loc)
    rst.loc <- crossprod(xt.y.loc, xt.x.loc.inv) %*% xt.y.loc
    x.loc = matrix(x[, mod.loc], nrow = n, ncol = p.loc)
    h <- sapply(seq(n), function(j) {
      crossprod(x.loc[j, ], xt.x.loc.inv) %*% x.loc[j, ]
    })
    r <- y - x.loc %*% hat.beta.loc

    rss.loc <- yt.y - rst.loc
    hat.sigmasq.loc <- rss.loc / n
    rel.rss.loc.marg <- (yt.y - rst.loc * n / (n + 1)) / beta.
    log.samp.at.mle <- -n / 2 * log(rss.loc)
    aic <- log.samp.at.mle * -2 + 2 * rank.loc
    bic <- log.samp.at.mle * -2 + rank.loc * log(n)
    tic <- log.samp.at.mle * -2 +
      2 * sum(r^2 * h) / hat.sigmasq.loc +
      0.5 * sum(r^4) / hat.sigmasq.loc^2 - 0.5

    log.marg <-
      -rank.loc / 2 * log(n + 1) - (n + alpha.) / 2 + log(1 + rel.rss.loc.marg)

    data.frame(i,  # keep track of the indices so we can match metrics to models
               p = p.loc,  # maybe worth looking at number of parameters used
               aic, bic, tic,
               log.samp.at.mle,
               log.marg)
  }) %>%
    dp$bind_rows()
}) %>%
  dp$bind_rows()

Sys.time() - t0
```

```
Time difference of 22.29439 mins
```

```r
out.df %>%
  dp$sample_frac(sample.frac) %>%
  ggplot() +
```

```r
geom_point(aes(x = aic, y = tic, colour = p)) +
scale_color_viridis()
```