STA 114: Statistics

Lab 6

Consider data X and a future variable of interest X^*, modeled as $X \sim f(x|\theta)$, $X^* \sim f^*(x^*|\theta)$, X and X^* independent, $\theta \in \Theta$. To predict X^* based on observations $X = x$, the classical approach suggests using the plug-in predictive distribution $f^*(x^*|\hat{\theta}_{\text{MLE}}(x))$ and the Bayesian approach leads to the posterior predictive distribution $f^*(x^*|x) = \int_\Theta f^*(x^*|\theta)\xi(\theta|x)d\theta$ under a prior $\xi(\theta)$ on θ. Recall that to simulate $x_1^*, \cdots, x_M^*$ from $f^*(x^*|x)$ it is equivalent to simulate $\theta_1, \cdots, \theta_M$ from $\xi(\theta|x)$ and then simulate x_i^* from $f^*(x^*|\theta = \theta_i)$. We will see this in action for a slightly complex model.

Let $X = (X_1, \cdots, X_n)$ and $X^* = X_{n+1}$ respectively stand for the annual hurricane counts for last n years and the next year. We model: $X_t \overset{\text{IND}}{\sim} \mathsf{Poisson}(\mu_t = e^{\alpha+\beta(t-1)})$, $t = 1, \cdots, n, n+1$, $(\alpha, \beta) \in [0, 5] \times [-1, 1]$. The log-likelihood function is given by (from 09/07 handout):

$$\ell_x(\alpha, \beta) = \text{const} - \sum_{t=1}^{n} \exp\{\alpha + \beta(t-1)\} + (\alpha - \beta)\sum_{t=1}^{n} x_t + \beta \sum_{t=1}^{n} t x_t.$$

Our observed data contains records from $n = 100$ years with $\sum_{t=1}^{n} x_t = 932$ and $\sum_{t=1}^{n} t x_t = 51884$.

The plug-in predictive pmf of X^*

You can obtain the mle of (α, β) by running the following code:

```
n <- 100; t <- 1:n; sum.x <- 932; sum.tx <- 51884
neg.log.lik <- function(par){
  alpha <- par[1]; beta <- par[2]; mu <- exp(alpha + beta * (t - 1))
  return(sum(mu) - (alpha - beta) * sum.x - beta * sum.tx)
}
o <- optim(c(0,0), neg.log.lik)
print(o)
alpha.hat <- o$par[1]; beta.hat <- o$par[2];
print(alpha.hat); print(beta.hat)
```

Task 1. Use the mle to make a plot of the plug-in predictive pmf of X^*. Plot over the range $x^* \in \{0, 1, \cdots, 50\}$

Task 2. Use the plug-in predictive and report the median and the central 95% credible interval for X^*.

Task 3. Calculate the probability of $X^* > 20$ under the plug-in predictive.

The posterior predictive of X^*

Now assign (α, β) a uniform prior pdf $\xi(\alpha, \beta) = \text{const.}$ over the range $[0, 5] \times [-1, 1]$. Then the posterior density $\xi(\alpha, \beta|x)$ satisfies

$$\log \xi(\alpha, \beta|x) = \text{const} + \ell_x(\alpha, \beta).$$

Our first task would be to plot the posterior density over a grid of points on the (α, β) space. We start with the following grid spanning the whole range

```
alpha.grid <- seq(0, 5, length = 101)
beta.grid <- seq(-1, 1, length = 101)
```

The code below computes $\xi(\alpha, \beta|x)$ (up to a multiplicative constant) at these grid points.

```
log.post <- Vectorize(function(a, b) return(-neg.log.lik(c(a, b))), c("a", "b"))
log.post.grid <- outer(alpha.grid, beta.grid, log.post)
post.grid <- exp(log.post.grid - max(log.post.grid))
```

And the code below makes a plot

```
image(alpha.grid, beta.grid, post.grid)
contour(alpha.grid, beta.grid, post.grid, add = TRUE)
```

TASK 4. Run the above code to produce an image+contour plot of the posterior density.

The posterior sits too tightly around its peak, which is same as the mle (why?), and the plot you get above is useless. So we would shrink the grid to a small region around the peak. We use the curvature of the log posterior (same as the curvature of $\ell_x(\alpha, \beta)$) to determine this region.

```
o <- optim(c(0,0), neg.log.lik, hessian = TRUE) ## hessian = 2nd deriv at optima
sd.par <- sqrt(diag(solve(o$hessian)))
sd.alpha <- sd.par[1]; sd.beta <- sd.par[2]
alpha.grid <- alpha.hat + 4 * sd.alpha * seq(-1, 1, length = 101)
beta.grid <- beta.hat + 4 * sd.beta * seq(-1, 1, length = 101)
```

TASK 5. Use the new grid to make an image+contour plot of the posterior density.

Next we turn attention to sampling from the posterior pdf $\xi(\alpha, \beta|x)$. This can be accomplished by discretization techniques we learned before. We will take the posterior values at our grid points and normalize to get a pmf over this grid and sample from the pmf. Here is a code for doing all this:

```
post.grid.long <- c(post.grid)
alpha.grid.long <- rep(alpha.grid, length(beta.grid))
beta.grid.long <- rep(beta.grid, each = length(alpha.grid))
grid.length <- length(post.grid.long)

M <- 1e4
ix.samp <- sample(1:grid.length, size = M, replace = TRUE, prob = post.grid)

alpha.samp <- alpha.grid.long[ix.samp]
beta.samp <- beta.grid.long[ix.samp]
```

TASK 6. Super-impose a plot of the sampled values of β against those of α over the image + contour plot you obtained before (use `points(alpha.samp, beta.samp, pch = ".")`)

TASK 7. Simulate a sample of values for X^* from $f^*(x^*|x)$.

TASK 8. Use this sample to approximate the posterior predictive median and the central 95% credible interval of $f^*(x^*|x)$.

TASK 9. Use this sample to compute $P(X^* > 20|X = x)$.