LECTURE 11: LIKELIHOOD BASED MODEL FITTING

CLASS INTRO

INTRO QUESTIONS

- Summarize how statistical simulation can be used to assess the quality of statistical estimators. More specifically, how is this process conducted?
- For Today:
  - Bayesian Hierarchical Models
  - Likelihood Based Model Fitting

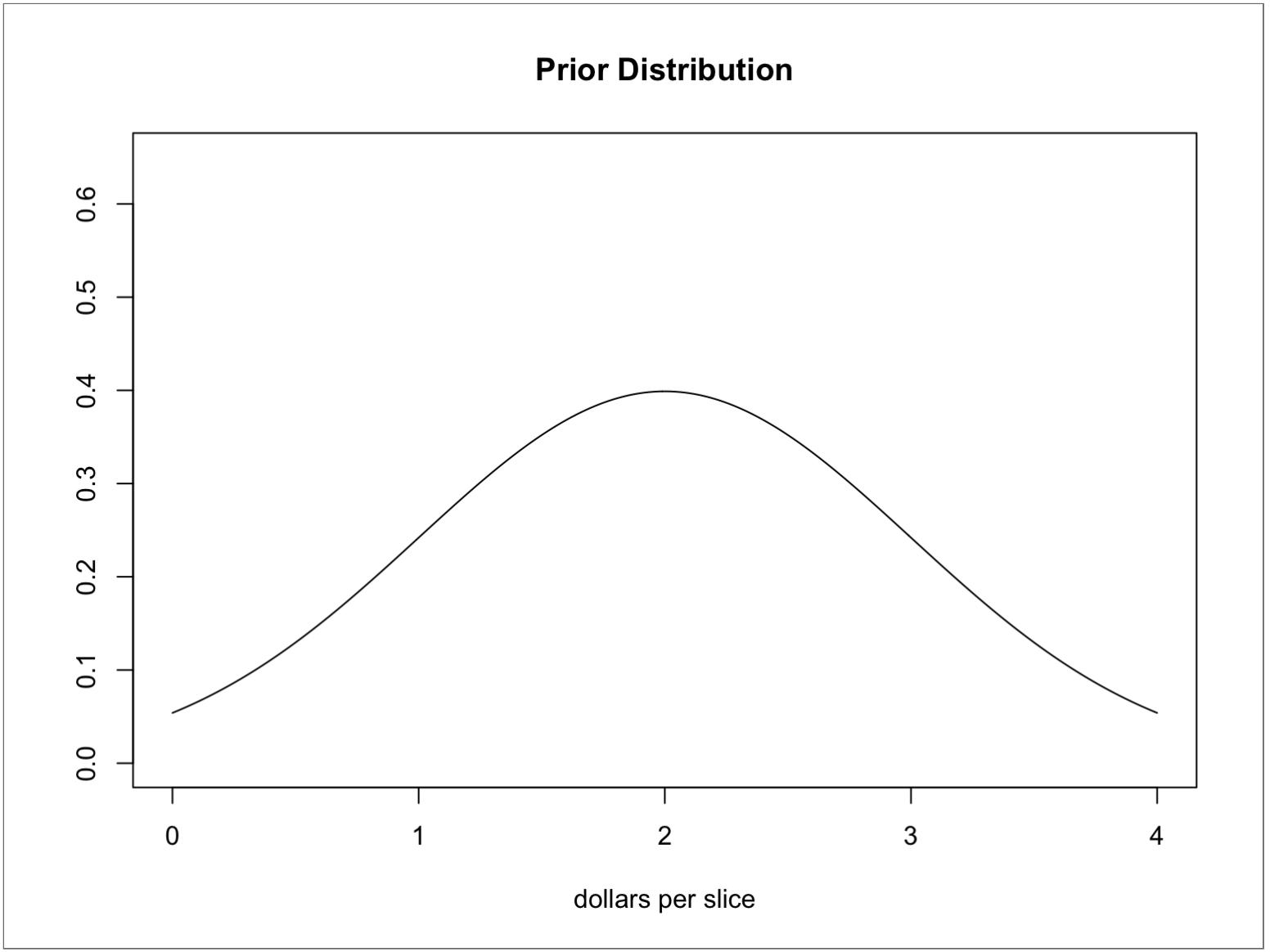
HIERARCHICAL MODELING FOR POINT REFERENCED DATA

MORE ABOUT BAYES

Bannerjee, Geland, and Carlin state, "Bayesian inferential paradigm offers potentially attractive advantages over the classical, frequentist statistical approach through

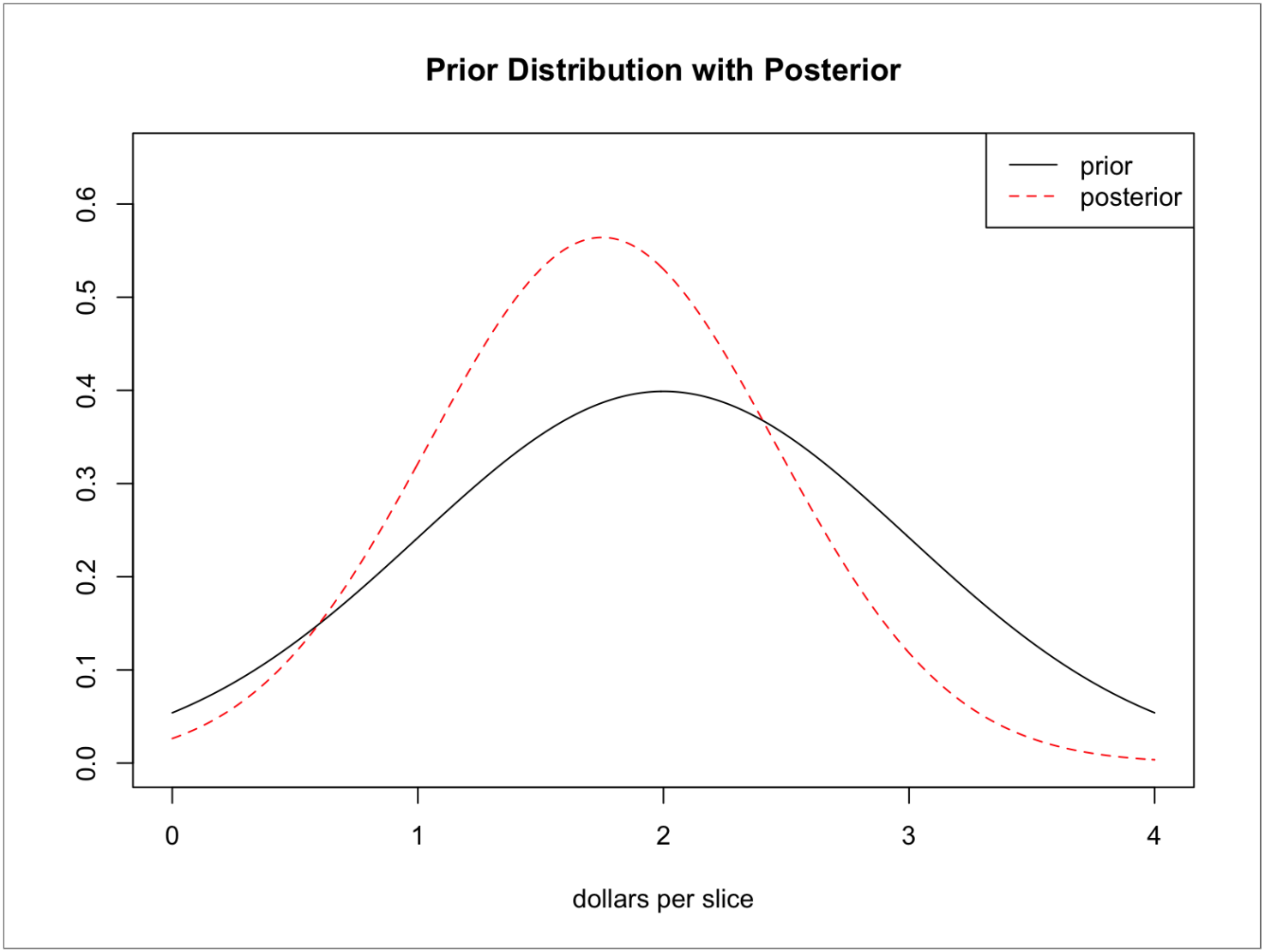
- its more philosophically sound foundation,
- its **unified approach to data analysis**,
- and its ability to incorporate prior opinion via the prior distribution.

BAYESIAN STATISTICS OVERVIEW: PRIOR DISTRIBUTION FOR PIZZA



BAYESIAN STATISTICS OVERVIEW: POSTERIOR DISTRIBUTION FOR PIZZA

- We observe a pizza with 8 slices sells for $12.



BAYESIAN STATISTICS OVERVIEW: HIERACHICAL POSTERIOR DISTRIBUTION

- Recall that $\boldsymbol{\lambda}$ are hyperparameters in the prior distribution $p(\boldsymbol{\theta}|\boldsymbol{\lambda})$.
- For instance, we might say that $\theta|\lambda \sim N(\lambda, 1)$
- Then we also need prior distributions (hyperpriors) for λ, $p(\lambda)$
- Then the posterior is

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y})p(\boldsymbol{\theta}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{\int\int \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y})p(\boldsymbol{\theta}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})d\boldsymbol{\theta}d\boldsymbol{\lambda}}$$

HIERARCHICAL MODEL

There are three levels of this (hierarchical) model

1. $p(\boldsymbol{y}|\boldsymbol{\theta})$ [data | process]
2. $p(\boldsymbol{\theta}|\boldsymbol{\lambda})$ [process | parameters]
3. $p(\boldsymbol{\lambda})$ [parameters]

STATIONARY SPATIAL PROCESS

- The model for a Gaussian process can be written as
  $Y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$,
  where $\mu(\boldsymbol{s}) = x(\boldsymbol{s})^t\boldsymbol{\beta}$ is the mean structure.
- Then the residual can be partitioned into two pieces: a spatial component $w(\boldsymbol{s})$ and a non-spatial component $\epsilon(\boldsymbol{s})$.
- We assume $w(\boldsymbol{s})$ are realizations from a Gaussian Process (GP) with mean zero.
- Then $\epsilon(\boldsymbol{s})$ are uncorrelated error terms.
- *Q:* how do $w(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$ relate to the partial sill, range, and nugget?

$\mathbf{w}(\boldsymbol{s})$ AND $\boldsymbol{\epsilon}(\boldsymbol{s})$

- The partial sill, σ^2, and the range, ϕ, are modeled with $w(\boldsymbol{s})$
- The nugget is contained in the $\epsilon(\boldsymbol{s})$ term.
- This framework assumes stationarity - in that the correlation is only a function of the separation between points.
- Furthermore, if the correlation is only a function of the distance between points, this is also isotropic.

MODEL SPECIFICATION

- Let $\Sigma = \sigma^2 H(\phi) + \tau^2 I$
- Define $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \tau^2, \phi)$
- Then the sampling model can be written as:
- $\boldsymbol{Y}|\boldsymbol{\theta} \sim N(X\boldsymbol{\beta}, \sigma^2 H(\phi) + \tau^2)$
- Given a (set of) prior distribution(s), $p(\theta)$, the posterior distribution of the parameters can be computed (more later) as $p(\theta|\boldsymbol{y})$.

MODEL SPECIFICATION AS HIERARCHICAL MODEL

- The model can be rewritten as
  $\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{W} \sim N(X\boldsymbol{\beta} + \boldsymbol{W}, \tau^2 I)$, [data | process, parameters]
  where $\boldsymbol{W} = (w(\boldsymbol{s_1}), \ldots, w(\boldsymbol{s_1}))^T$ is a vector of spatial random effects.
- The second-stage, or the process, is
  $\boldsymbol{W}|\sigma^2, \phi \sim N(\boldsymbol{0}, \sigma^2 H(\phi))$ [process | parameters]
- The third level is the prior specification: $p(\theta)$ [parameters]

WHAT ABOUT THOSE PRIOR DISTRIBUTIONS??

Next, prior distributions are necessary for

- $\boldsymbol{\beta}$
- σ^2
- ϕ
- τ^2

PRIOR SELECTION

- I generally advocate, fairly objective prior beliefs. In other words, priors that are not highly influential on the posterior distribution.
- Conjugate (or semi-conjugate) priors, in general, are useful as they make computation more efficient.
- In the Mate'rn class of covariance functions, including the exponential, the range and partial sill parameters are not individually identifiable. This does not impact the Kriging result, but rather just inferences about the parameters. The textbook suggests a very informative prior on ϕ and a vague prior on σ^2.

BAYESIAN COMPUTING: MCMC

- Many Bayesian algorithms use Markov Chain Monte Carlo (MCMC) to estimate the joint posterior distribution of the parameters, $p(\boldsymbol{\beta}, \sigma^2, \phi, \tau^2 | y)$ in this case.
- **Goal**: Describe MCMC to a classmate that has not yet taken a Bayesian statistics course.
- The end result is a joint posterior distribution that represents the uncertainty in the parameter estimates.

SAMPLING IS INTEGRATING

- The joint posterior can be written as

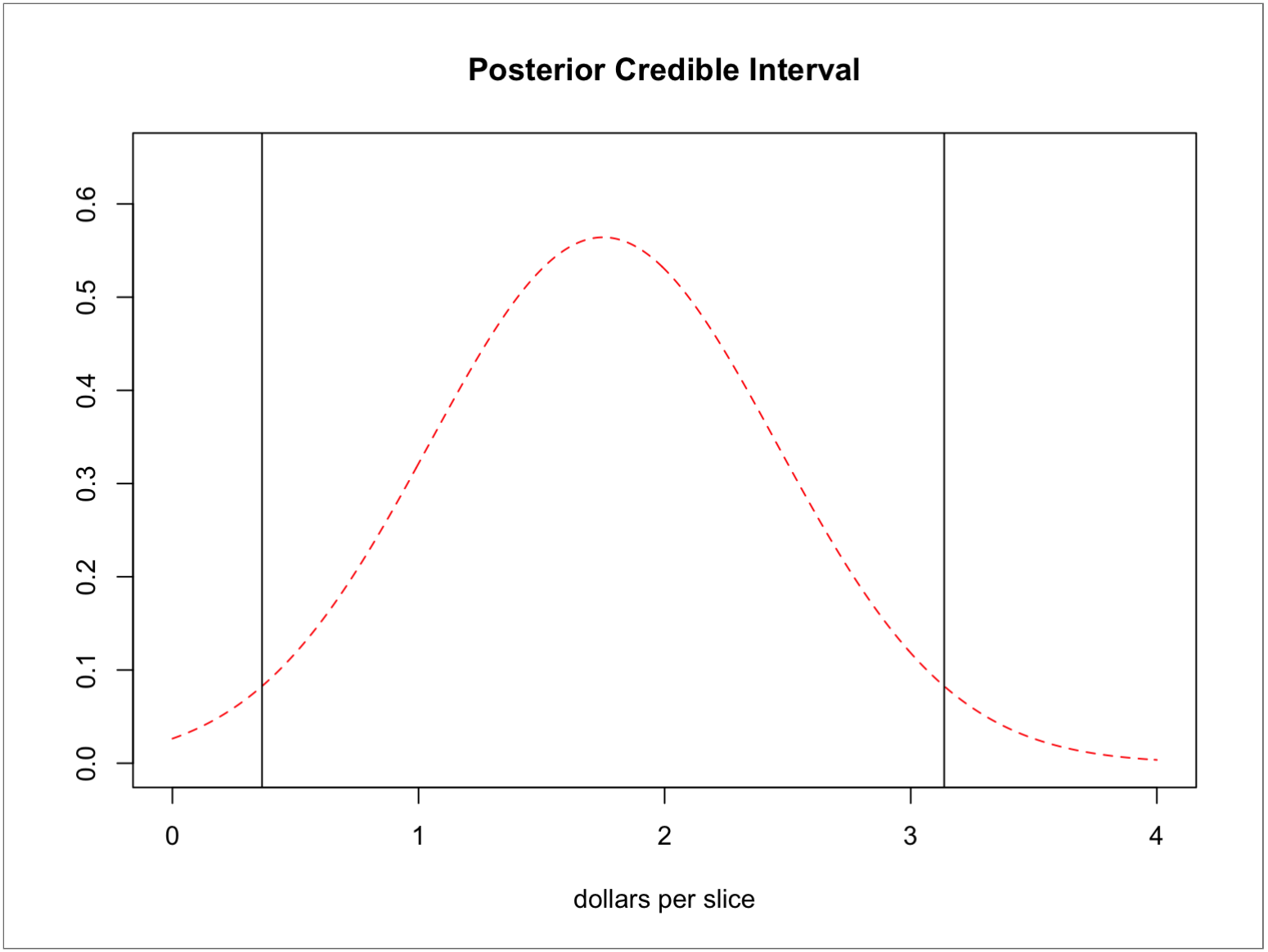
$$p(\boldsymbol{\beta}, \sigma^2, \phi, \tau^2 | y) = \frac{\mathcal{L}(\boldsymbol{\beta}, \sigma^2, \phi, \tau^2 | \boldsymbol{y}) p(\boldsymbol{\beta}, \sigma^2, \phi, \tau^2)}{\int\int\int\int \mathcal{L}(\boldsymbol{\beta}, \sigma^2, \phi, \tau^2 | \boldsymbol{y}) p(\boldsymbol{\beta}, \sigma^2, \phi, \tau^2) d\boldsymbol{\beta} d\sigma^2 d\phi d\tau^2}$$

- The integration is conducted by taking MCMC samples.
- Similarly by taking samples, we can obtain marginal posterior distributions, such as,

$$p(\boldsymbol{\beta}|y) = \int\int\int p(\boldsymbol{\beta}, \sigma^2, \phi, \tau^2 | y) d\sigma^2 d\phi d\tau^2$$

INTERVAL ESTIMATION

- Posterior estimates are typically reported using *credible intervals*.


Posterior Credible Interval
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
2
3
4
dollars per slice

PREDICTIVE DISTRIBUTIONS

- The Bayesian prediction distribution for can be written as

$$p(y_0 | \boldsymbol{y}, X, x_0) = \int p(y_0, \boldsymbol{\theta} | \boldsymbol{y}, X, x_0) d\boldsymbol{\theta}$$

- Mathematically, this is similar to the Kriging predictions we saw earlier, but rather than conditioning on $\boldsymbol{\theta}$, the parameters are integrated out.
- The implication of this is that the uncertainty in the parameter estimates are captured and propogated in the posterior predictive distribution.

LIKELIHOOD BASED MODEL FITTING

VARIOGRAM BASED MODEL FITTING

- Up until now, we have used a least-squares approach with the variogram to estimate the covariance parameters.
- How would you do this if there were covariates that could be used to explain the process?
- Use the residuals from a linear model.

`KRIGE.BAYES()` DEMO

- For this demonstration we will explore the `krige.bayes()` function in R using a modified script from the function description. With this exploration, answer the following questions.

1. What does the `grf()` function do?
2. Explain the parameters in the `prior.control()` section.
3. Describe the output from `hist(ex.bayes)`.
4. What are the four figures generated from the `image()` function?

CODE

```
set.seed(02132019)
# generating a simulated data-set
ex.data <- grf(75, cov.pars=c(10, .15), cov.model="exponential", nugget = 1)
#
# defining the grid of prediction locations:
ex.grid <- as.matrix(expand.grid(seq(0,1,l=15), seq(0,1,l=15)))
#
# computing posterior and predictive distributions
# (warning: the next command can be time demanding)
ex.bayes <- krige.bayes(ex.data, loc=ex.grid,
                        model = model.control(cov.m="exponential"),
                        prior = prior.control(beta.prior = 'flat',
                                              sigmasq.prior = 'reciprocal',
                                              phi.discrete=seq(0, 0.7, l=25),
                                              phi.prior="uniform",
                                              tausq.rel.discrete = seq(0, 1,
  l=25),
                                              tausq.rel.prior = 'uniform'))

# Plot histograms with samples from the posterior
par(mfrow=c(4,1))
hist(ex.bayes)
par(mfrow=c(1,1))
```

`SPLM()` DEMO

- Another option for fitting Bayesian spatial models is the `spLM()` function in the `spBayes` package. Using the code on the next slide, answer the following questions.

1. What is `w`?
2. What does the `tuning` argument in `spLM()` control?
3. What does the following code return `summary(m.1$p.beta.recover.samples)$quantiles`?
4. Describe the final figure generated by this code.

CODE

```
rmvn <- function(n, mu=0, V = matrix(1)){
  p <- length(mu)
  if(any(is.na(match(dim(V),p))))
    stop("Dimension problem!")
  D <- chol(V)
  t(matrix(rnorm(n*p), ncol=p)%*%D + rep(mu,rep(n,p)))
}

n <- 100
coords <- cbind(runif(n,0,1), runif(n,0,1))
X <- as.matrix(cbind(1, rnorm(n)))

B <- as.matrix(c(1,5))
p <- length(B)

sigma.sq <- 2
tau.sq <- 0.1
phi <- 3/0.5

D <- as.matrix(dist(coords))
R <- exp(-phi*D)
w <- rmvn(1, rep(0,n), sigma.sq*R)
y <- rnorm(n, X%*%B + w, sqrt(tau.sq))
```

OTHER MODELING OPTIONS

- JAGS is another option for fitting general Bayesian models.
- Additionally, these models can also be implemented from scratch.