

LECTURE 10: POINT LEVEL MODELS - MODEL FITTING, CONT..

CLASS INTRO

INTRO QUESTIONS

- Discuss how the spatial range could be extended to \mathcal{R}^2 rather than \mathcal{R}^1 in the Gaussian Process setting. What changes in this situation?
- For Today:
 - More Model Fitting
 - Hierarchical Models Intro

UNIVERSAL KRIGING

- When covariate information is available for inclusion in the analysis, this is often referred to as *universal kriging*
- Now we have
$$Y = X\beta + \epsilon, \quad \text{where } \epsilon \sim N(\mathbf{0}, \Sigma)$$
- The conditional distributions are very similar to what we have derived above, watch for HW question.
- In each case, kriging or universal kriging, it is still necessary to estimate the following parameters: σ^2 , τ^2 , ϕ , and μ or β .
- This can be done with least-squares methods or in a Bayesian framework.

PARAMETER ESTIMATION

ABOUT THOSE OTHER PARAMETERS

We still need

- to choose an appropriate covariance function (or semivariogram)
- and estimate parameters in that function

FUNCTIONS IN R FOR 2D KRIGING

- The `krige` function in `gstat` contains a function for kriging ; however, again this requires a known variogram.

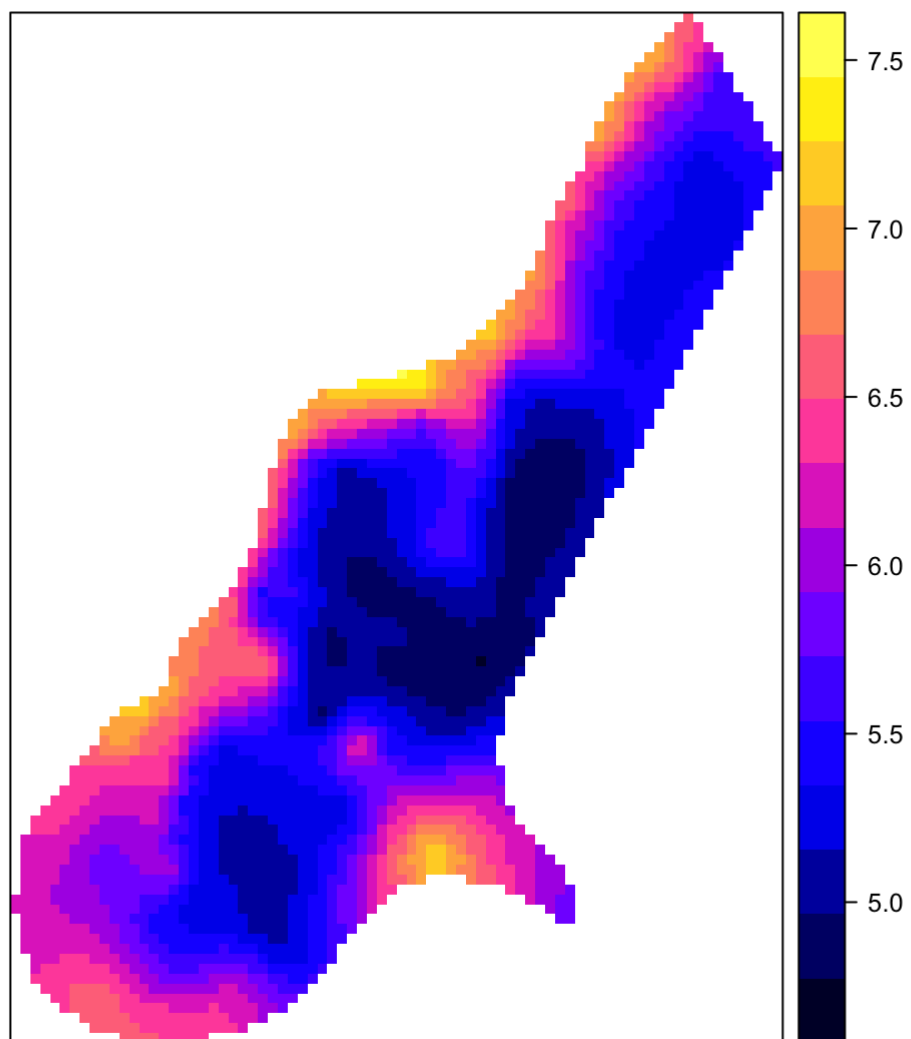
```
data(meuse)
coordinates(meuse) = ~x +y
data(meuse.grid)
gridded(meuse.grid) = ~x + y
m <- vgm(.59, "Sph", 874, .04)
# ordinary kriging:
x <- krige(log(zinc)~1, meuse, meuse.grid, model = m)
spplot(x["var1.pred"], main = "ordinary kriging predictions")
```

FUNCTIONS IN R FOR 2D KRIGING

- The `krige` function in `gstat` contains a function for kriging ; however, again this requires a known variogram.

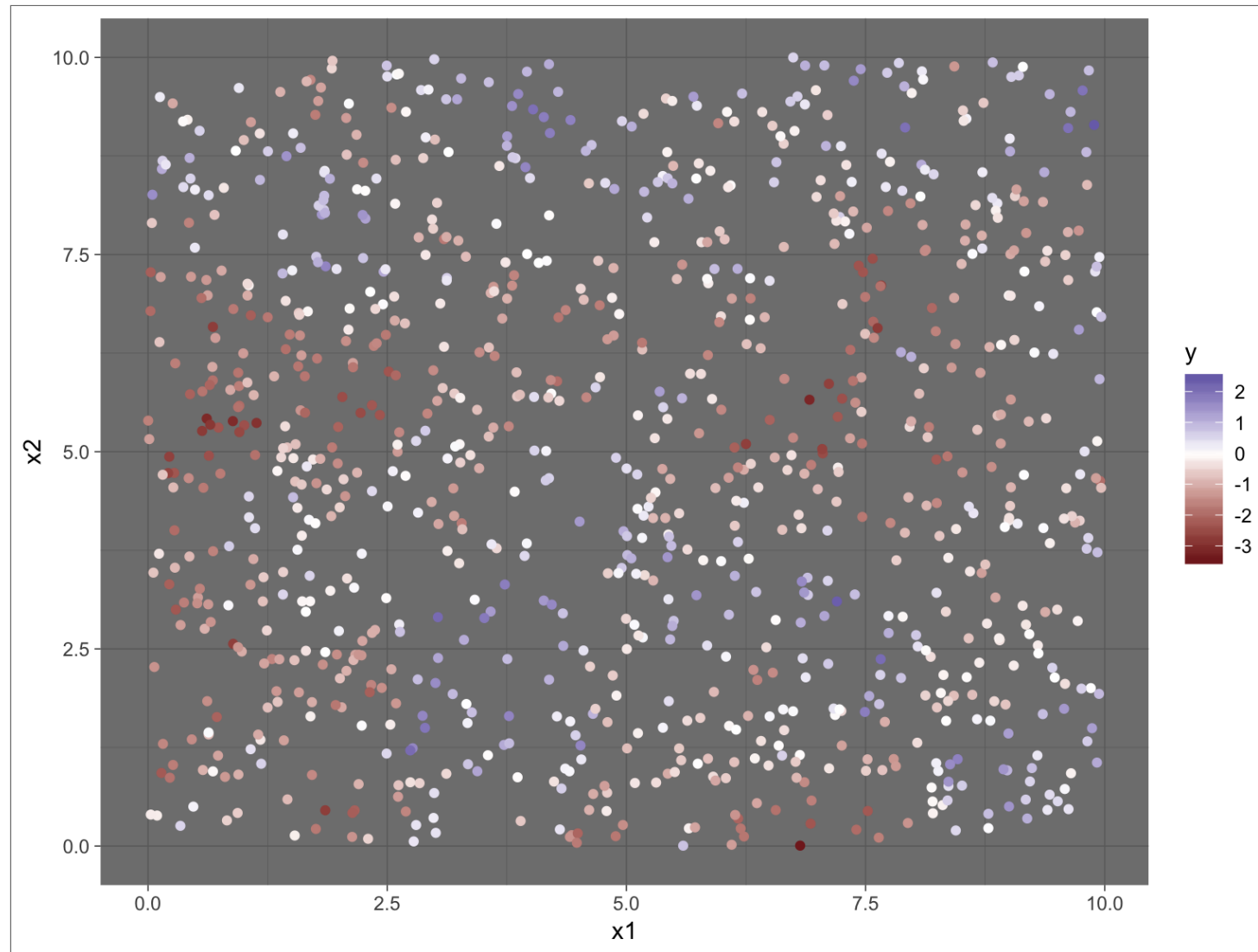
```
## [using ordinary kriging]
```


ordinary kriging predictions

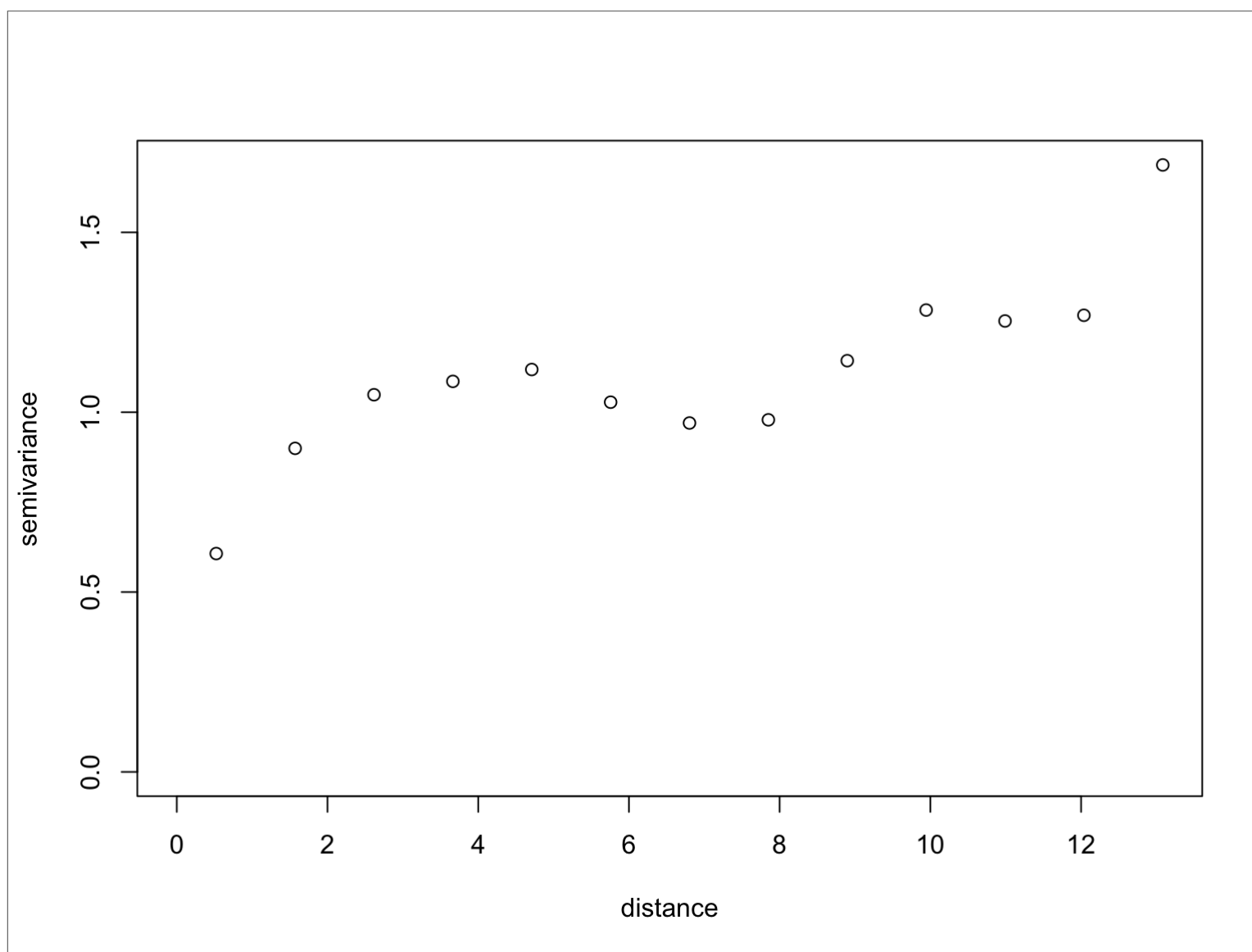


2D SIMULATION

- Using the simulated spatial process, plot an empirical variogram.



VARIOGRAM FITTING: OPTIMIZATION TO EMPIRICAL VARIOGRAM



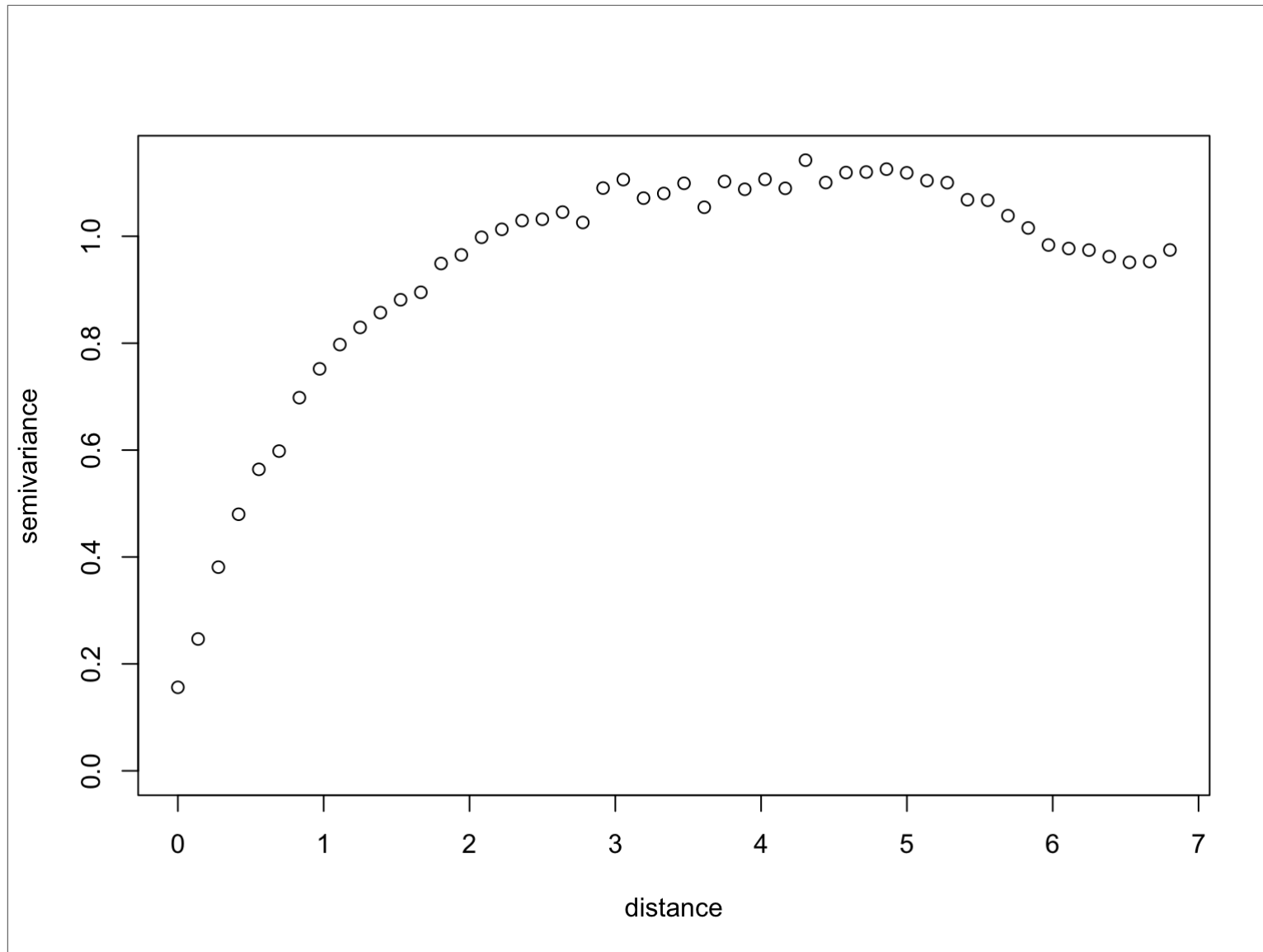
```
## variofit: covariance model used is exponential
## variofit: weights used: npairs
## variofit: minimisation function used: optim
```

```
## Warning in variofit(emp.vario, cov.model = "exponential"): initial values
## not provided - running the default search
```

```
## variofit: searching for best initial value ... selected values:
##           sigmasq phi   tausq kappa
## initial.value "0.84"  "0"   "0.17" "0.5"
## status        "est"   "est" "est"  "fix"
## loss value: 6873.17015373233
```

```
## variofit: model parameters estimated by WLS (weighted least squares):
## covariance model is: exponential
## parameter estimates:
##   tausq sigmasq      phi
## 0.1797 0.8546 0.0000
## Practical Range with cor=0.05 for asymptotic range: 0.0001159668
##
## variofit: minimised weighted sum of squares = 6633.819
```

VARIOGRAM FITTING: OPTIMIZATION TO EMPIRICAL VARIOGRAM



```
## variofit: covariance model used is exponential
## variofit: weights used: npairs
## variofit: minimisation function used: optim
```

```
## Warning in variofit(emp.vario2, cov.model = "exponential"): initial values
## not provided - running the default search
```

```
## variofit: searching for best initial value ... selected values:
##           sigmasq phi      tausq kappa
## initial.value "1.14"  "1.09"  "0"    "0.5"
## status        "est"   "est"   "est"  "fix"
## loss value: 2384.34387917209
```

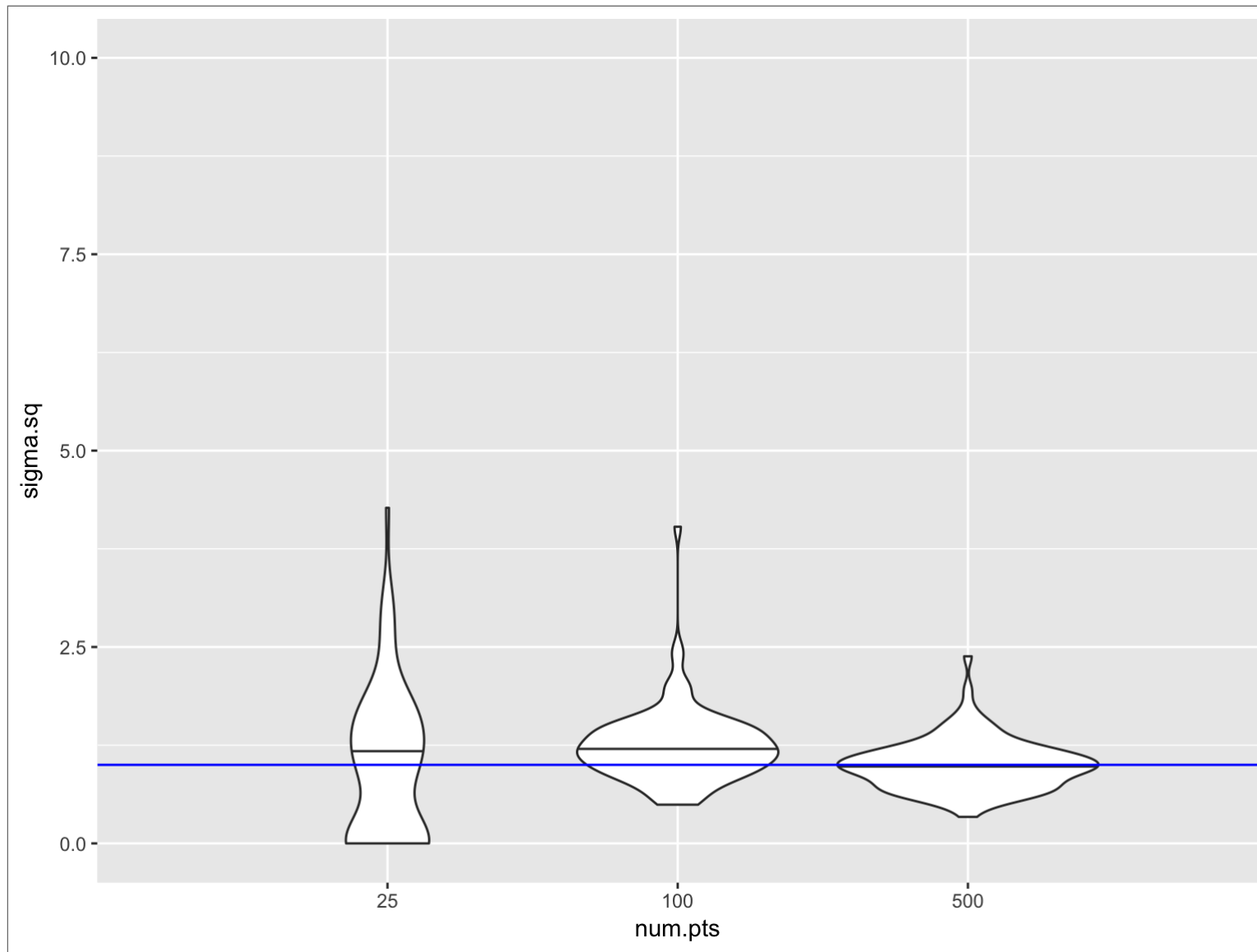
```
## variofit: model parameters estimated by WLS (weighted least squares):
## covariance model is: exponential
## parameter estimates:
##   tausq sigmasq      phi
## 0.0189 1.0468 0.7845
## Practical Range with cor=0.05 for asymptotic range: 2.350179
##
## variofit: minimised weighted sum of squares = 1035.009
```

VARIOGRAM FITTING: EXERCISE

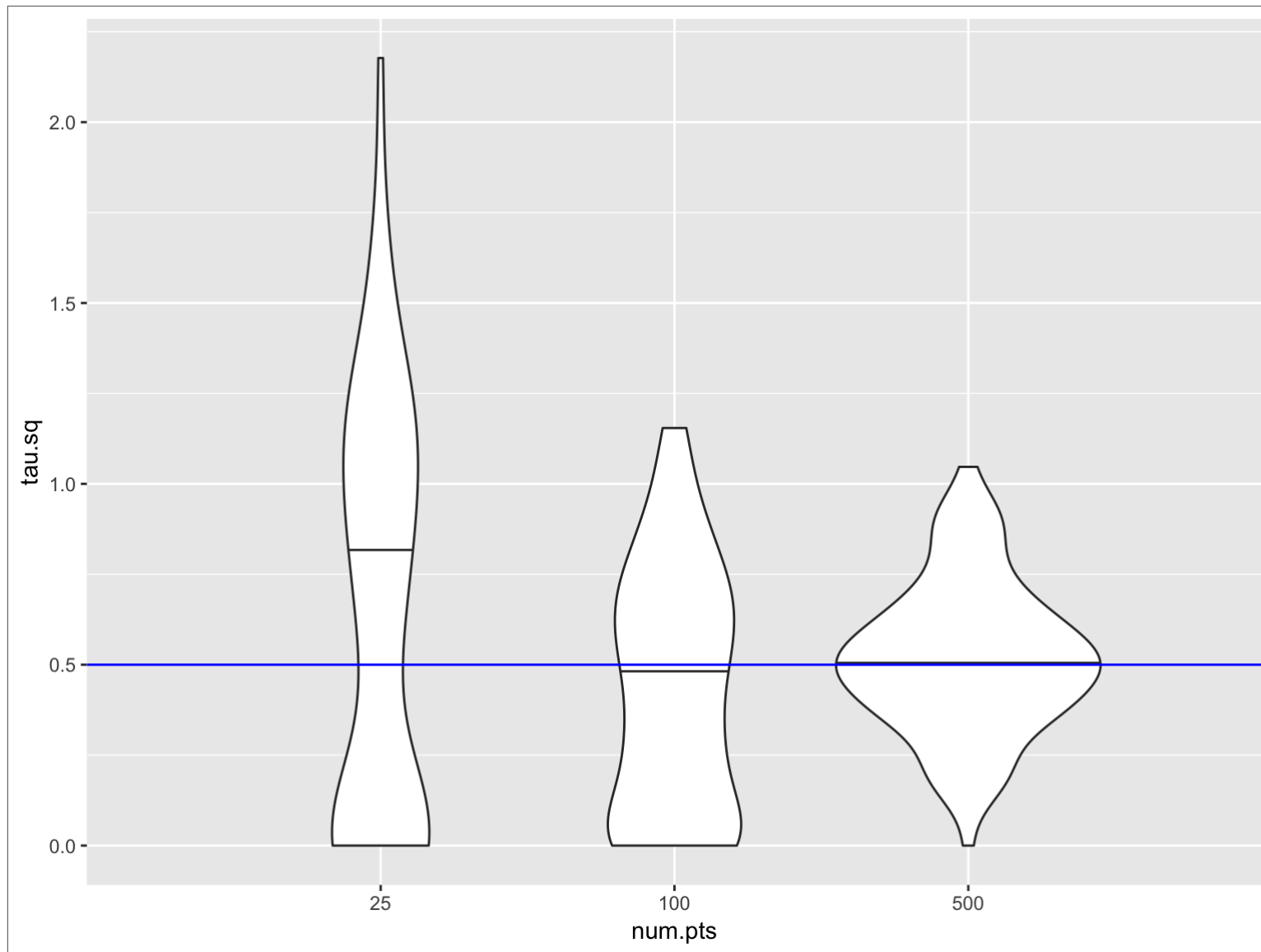
- Write code to simulate spatial processes with an exponential covariance function and evaluate the fitted values for the parameters.
- Vary the following parameters: number of samples (25, 100, 500) and replicate each 10 times. Discuss how close the fitted samples are to the true values.

VARIOGRAM FITTING: SOLUTION σ^2

```
## Warning: Removed 35 rows containing non-finite values (stat_ydensity).
```

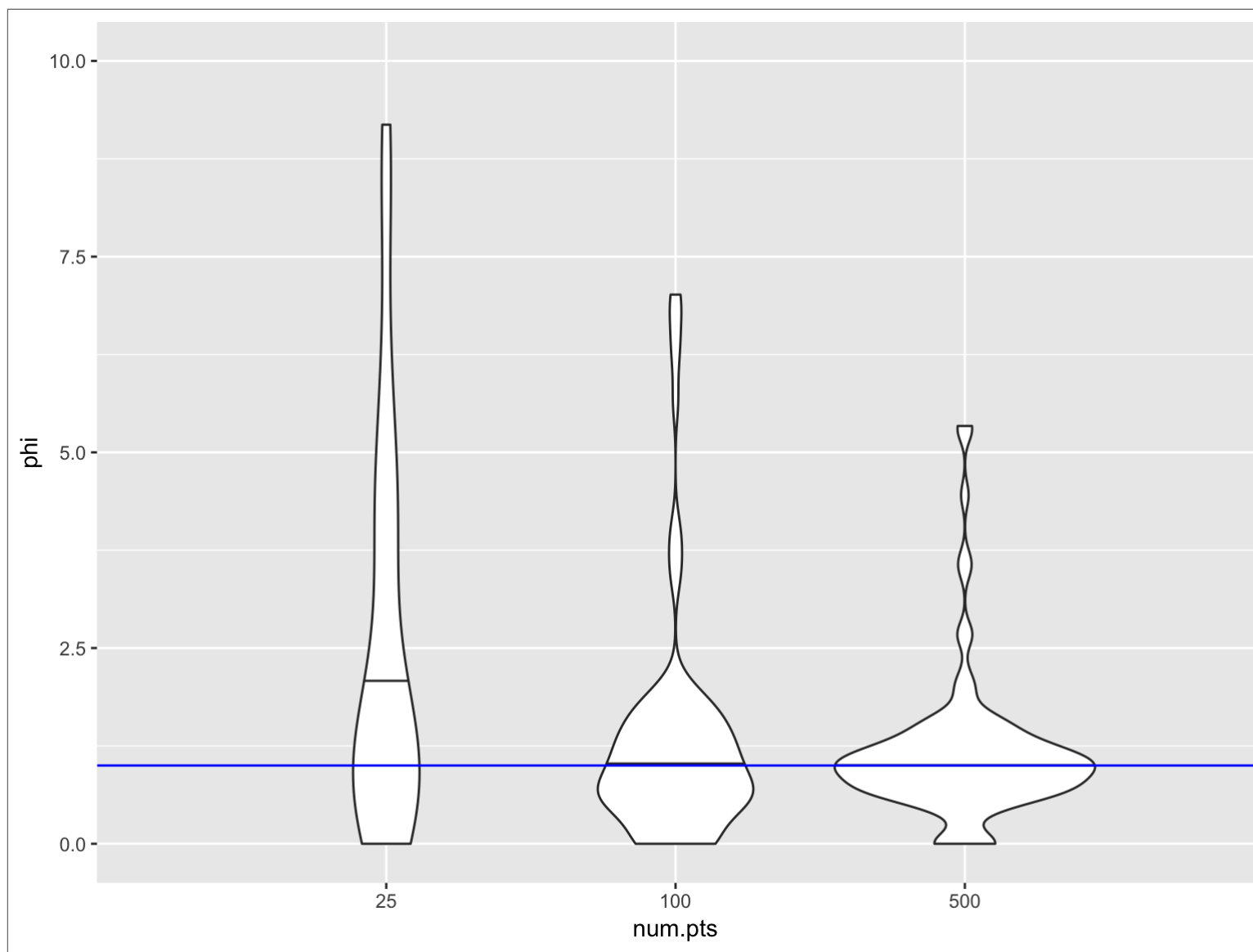


VARIOGRAM FITTING: SOLUTION τ^2



VARIOGRAM FITTING: SOLUTION ϕ

```
## Warning: Removed 43 rows containing non-finite values (stat_ydensity).
```



HIERARCHICAL MODELING FOR POINT REFERENCED DATA

SPATIAL PROCESS THEORY

- Specifying a spatial process is concerned with a collection of random variables $Y(s) : s \in \mathcal{R}^2$
- Recall, $s \in \mathcal{R}^2$ requires an infinite dimensional distribution.
- This is typically achieved using a Gaussian process, which enables use of a multivariate normal distribution and only requires specifying the mean and covariance function.

GAUSSIAN PROCESS

- In practice, the process is only observed at a finite number of locations, $\{s_1, \dots, s_n\}$
- Using $\{Y(s_1), \dots, Y(s_n)\}$ the goal is to learn the mean and covariance structure of the latent spatial process using observed data
- *and* make predictions of $Y(s_0)$
- A common statistical modeling approach uses Bayesian hierarchical models

CONNECTIONS AND ASYMPTOTICS

- This is quite similar to dynamic linear models where there is a latent state-space specification.
- Asymptotics focuses on the convergence of estimators with large samples. In spatial statistics, asymptotics are focused on spatial infill. That is, considering the properties of the estimators as the distance between the points goes to zero.

BAYESIAN STATISTICS OVERVIEW: SAMPLING MODEL

- Bayesian statistics assumes that the unknown parameters are a random variable - in addition to observed data.
- As with classical statistical techniques, fitting a model requires specifying the statistical distribution for the observed data $\mathbf{y} = \{y_1, \dots, y_n\}$ given parameters $\boldsymbol{\theta}$, denoted as $p(\mathbf{y}|\boldsymbol{\theta})$.
- $p(\mathbf{y}|\boldsymbol{\theta})$ is often referred to as a sampling model.
- For instance, with a regression model,
 $y|\boldsymbol{\beta}, \mathbf{x} \sim ??$

BAYESIAN STATISTICS OVERVIEW: SAMPLING MODEL

- With a regression model, typically the residuals are assumed to follow a normal distribution.

$$Y_i = \beta^T x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Hence,

$$y|\beta, x \sim N(X\beta, \sigma^2 I)$$

- In regression, the goal is to make inferences about the parameters $\theta = \{\beta, \sigma^2\}$

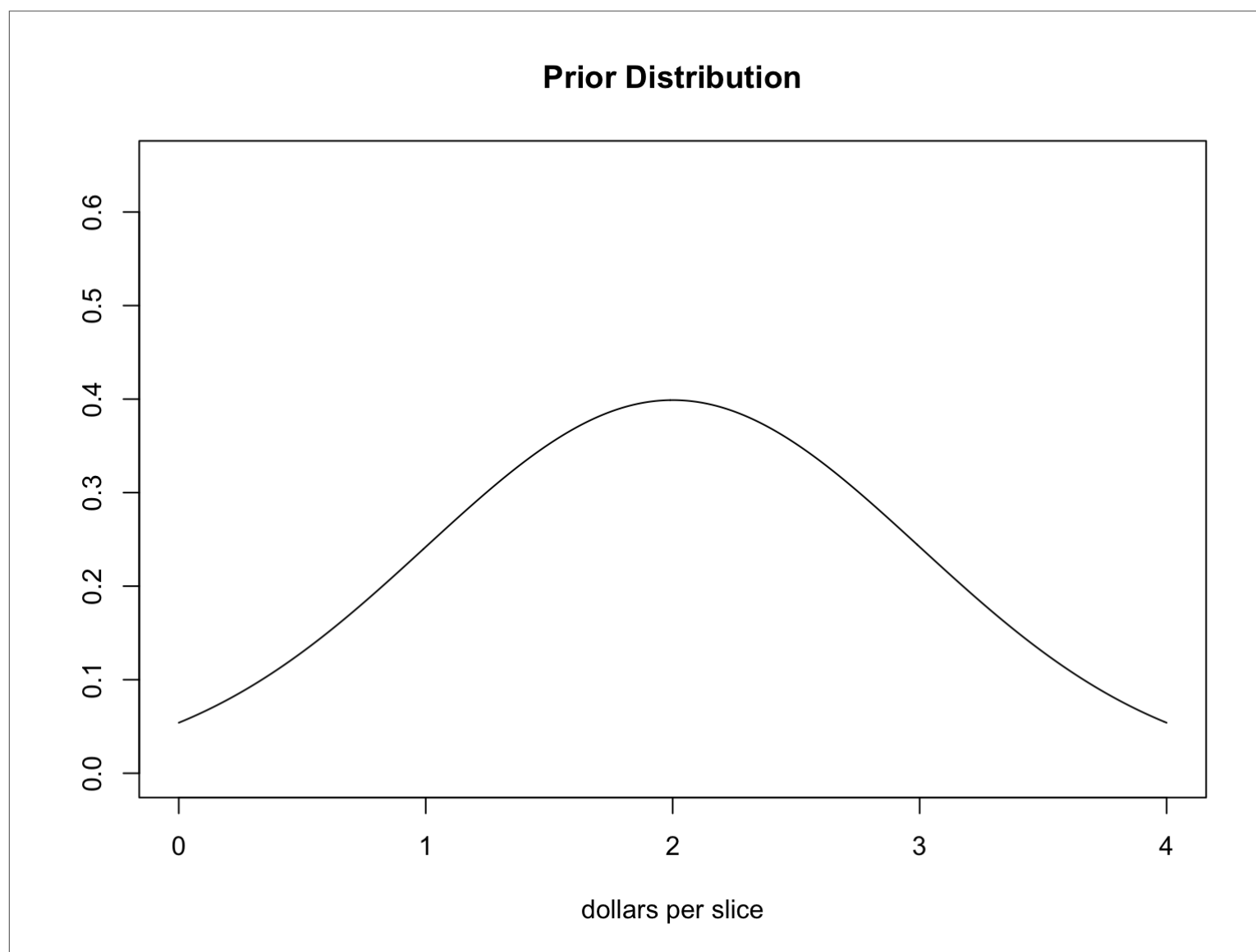
BAYESIAN STATISTICS OVERVIEW: PRIOR DISTRIBUTION

- Inference in Bayesian statistics is distributional, rather than a single point.
- Distributional beliefs are iteratively updated in the presence of new data
- A prior distribution, $p(\theta|\lambda)$ is the belief about the parameters(θ) before collecting data.
- A set of hyperparameters, λ can be used to define the prior distribution.

BAYESIAN STATISTICS OVERVIEW: PRIOR DISTRIBUTION

- Assume we are interested in learning the relationship between the cost of a pizza in Bozeman and the number of slices in that pizza.
- Write out and specify a sampling model for this regression problem.
- Sketch and/or define a prior distribution for your parameters in the sampling model.

BAYESIAN STATISTICS OVERVIEW: PRIOR DISTRIBUTION FOR PIZZA



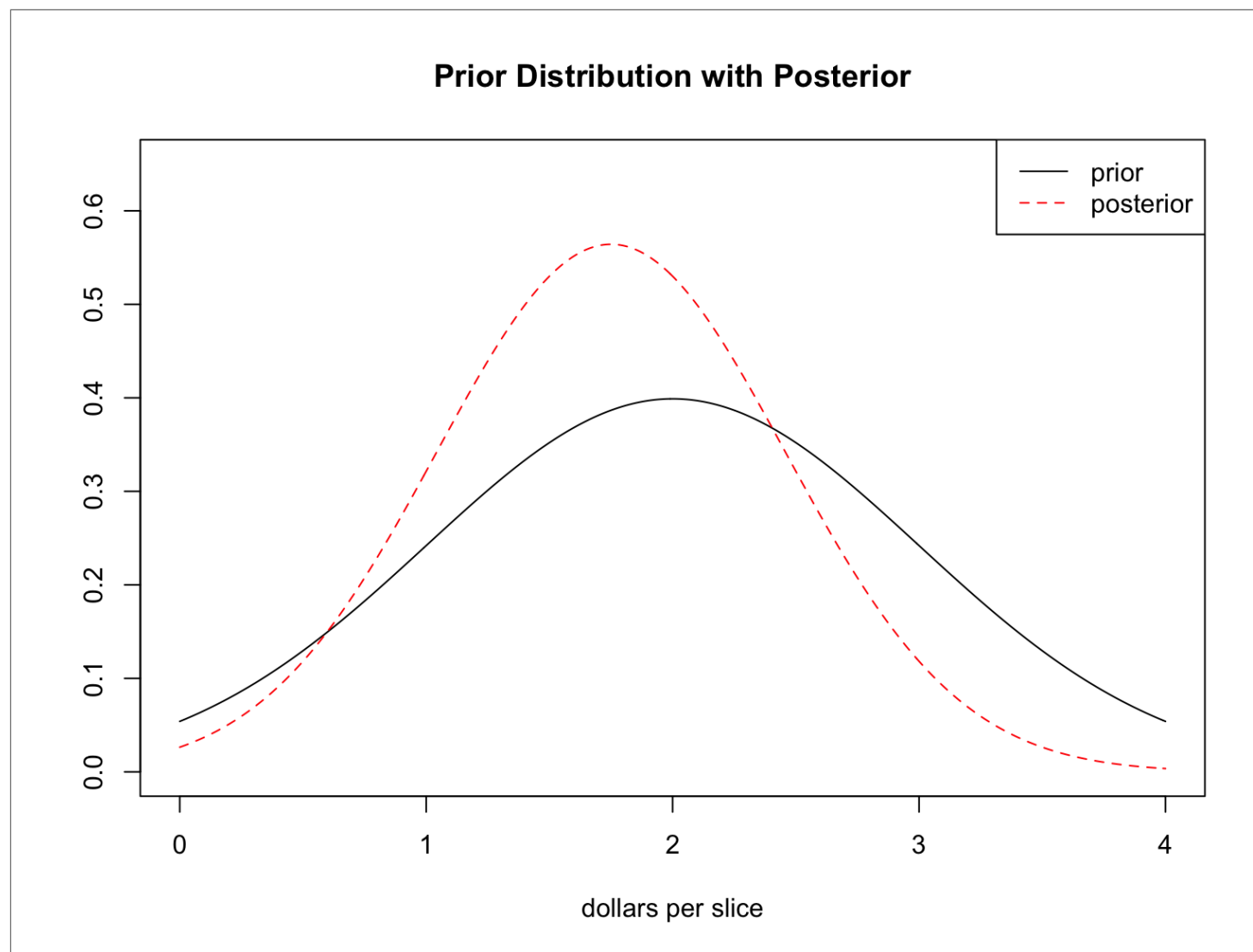
BAYESIAN STATISTICS OVERVIEW: POSTERIOR DISTRIBUTION

- The combination of the prior distribution, which is then updated after observing data, results in the posterior distribution.
- Inference about the parameters is based on the posterior distribution.
- Ignoring λ for now, the posterior distribution is defined as

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

BAYESIAN STATISTICS OVERVIEW: POSTERIOR DISTRIBUTION FOR PIZZA

- We observe a data point where a pizza with 8 slices sells for \$12.



BAYESIAN STATISTICS OVERVIEW: HIERACHICAL POSTERIOR DISTRIBUTION

- Recall that λ are hyperparameters in the prior distribution $p(\boldsymbol{\theta}|\lambda)$.
- 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

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

HIERARCHICAL MODEL

There are three levels of this (hierarchical) model

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

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.

STATIONARY SPATIAL PROCESS

-The model for a Gaussian process can be written as

$$Y(s) = \mu(s) + w(s) + \epsilon(s),$$

where $\mu(s) = x(s)^t \boldsymbol{\beta}$ is the mean structure. Then the residual can be partitioned into two pieces: a spatial component $w(s)$ and a non-spatial component $\epsilon(s)$.

- We assume $w(s)$ are realizations from a Gaussian Process (GP) with mean zero that models residual spatial structure.
- Then $\epsilon(s)$ are uncorrelated error terms.
- *Q*: how do $w(s) + \epsilon(s)$ relate to the partial sill, range, and nugget?

$w(s)$ AND $\epsilon(s)$

- The partial sill, σ^2 , and the range, ϕ , are modeled with $w(s)$
- The nugget is contained in the $\epsilon(s)$ term.
- This model assumes a stationary model - 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 $\theta = (\beta, \sigma^2, \tau^2, \phi)$
- Then the sampling model can be written as:
- $Y|\theta \sim N(X\beta, \sigma^2 H(\phi) + \tau^2)$
- Given a (set of) prior distribution(s), $p(\theta)$, the posterior distribution of the parameters can be computed (more later) as $p(\theta|y)$.

MODEL SPECIFICATION AS HIERARCHICAL MODEL

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

WHAT ABOUT THOSE PRIOR DISTRIBUTIONS??

Next, prior distributions are necessary for

- β
- σ^2
- ϕ
- τ^2