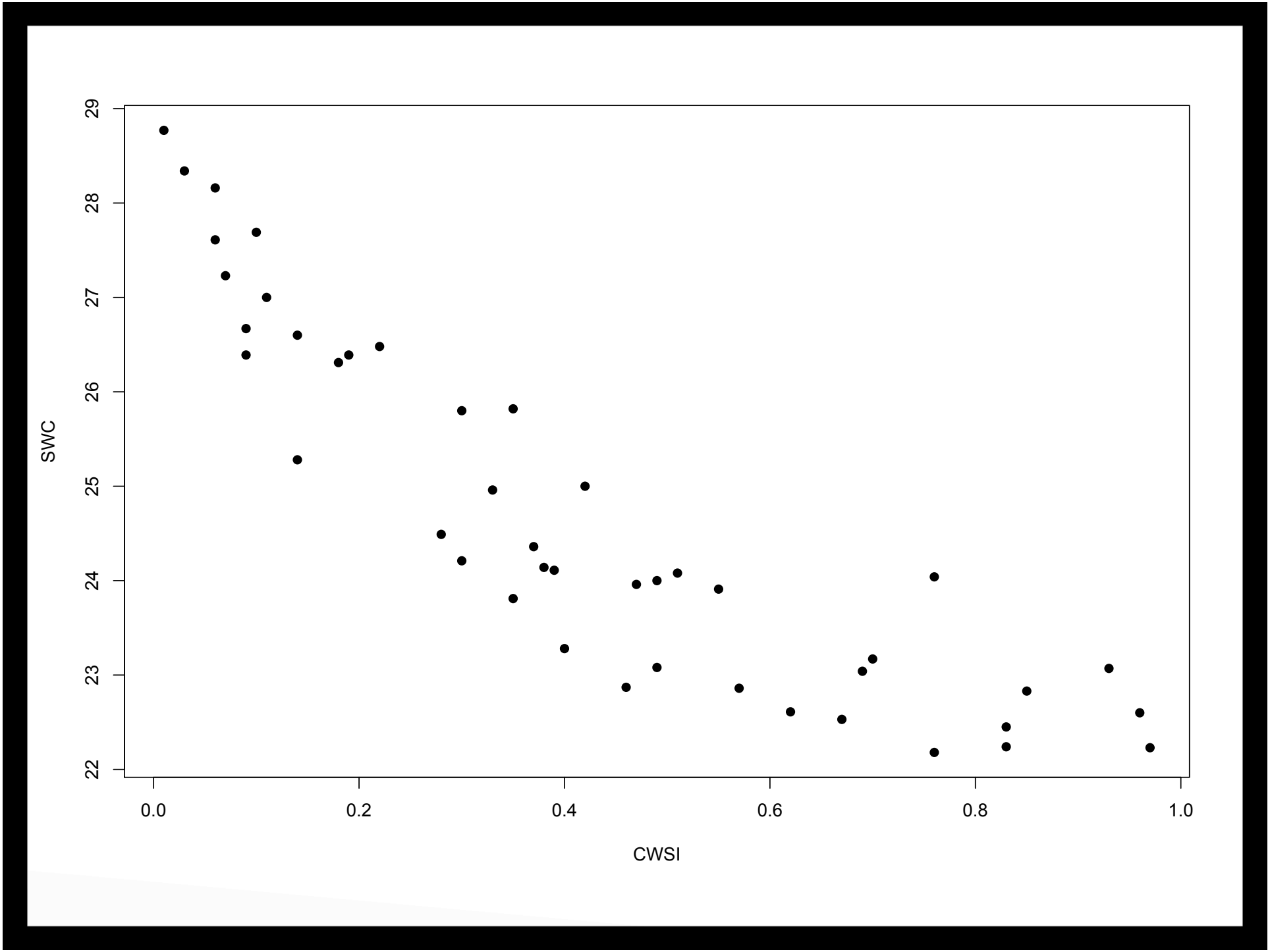
Gaussian Process Regression for the Agriculture Analysis


SWC
29
28
27
26
25
24
23
22
0.0
0.2
0.4
0.6
0.8
1.0
CWSI

Agriculture Data

What is our goal with the Agriculture data?

1. Establish relationship between CWSI and SWC. Want to know how well CWSI explains SWC.
2. We want to predict SWC.

Why is this goal important?

- Use CWSI to establish how much water is in the soil and know how much water to add to crops. This helps in better managing water resources.

Agriculture Data

What are the challenges with the Agriculture data?

1. Non-linear relationship.

How can we deal with non-linearity?

- Transformations
- Polynomial Regression
- Splines
- Wavelets
- Kernel Smoothing
- Local Regression
- Smoothing Splines
- Gaussian Process Regression

Review of MVN Distribution

Let $\mathbf{Y} = (y_1, \ldots, y_P)'$. If $\mathbf{Y}$ follows a multivariate normal (Gaussian) distribution then,

$$\mathbf{Y} \sim \mathcal{N}_P(\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y)$$

$$\Rightarrow f_{\mathbf{Y}}(\mathbf{y}) = \left(\frac{1}{2\pi}\right)^{P/2} \frac{1}{|\boldsymbol{\Sigma}_Y|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}_Y^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}$$

where, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_P)'$ is the mean vector and $\boldsymbol{\Sigma}_X$ is the covariance matrix.

Review of MVN Distribution

Partition,

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \ \boldsymbol{\Sigma}_Y = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}'_{12} & \boldsymbol{\Sigma}_2 \end{pmatrix}$$

The marginal distribution of $\mathbf{Y}_1$ is $\mathbf{Y}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$.

The conditional distribution of $\mathbf{Y}_1 \mid \mathbf{Y}_2$ is

$$\mathbf{Y}_1 \mid \mathbf{Y}_2 \sim \mathcal{N}\left(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}\right)$$

where

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1}\left(\mathbf{Y}_2 - \boldsymbol{\mu}_2\right)$$

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}'_{12}$$

Review of MVN Distribution

How to draw from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

1. Calculate Cholesky Decomposition $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}'$
2. Draw $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
3. Set $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{L}\mathbf{Z}$

```
mvn.draw <- mu+t(chol(Sigma))%*%rnorm(p)
```

Can you show:

$$\mathbb{E}(\mathbf{Y}) = \boldsymbol{\mu}$$

$$\mathrm{V}(\mathbf{Y}) = \boldsymbol{\Sigma}$$

Stochastic Processes

Stochastic Process: A (potentially infinite) collection of random variables over a domain $\mathcal{T}$.

- A "realization" or "observation" of a stochastic process is a finite collection of random variables at points $t_1, \ldots, t_N \in \mathcal{T}$.
- Examples:
  - Time Series: $\mathcal{T}$ is the temporal domain and we observe $Y(t_1), \ldots, Y(t_N)$ for N temporal "locations" $t_1, \ldots, t_N$.
  - Spatial: $\mathcal{T} = \mathcal{D}$ (or $\mathcal{S}$) is a spatial domain (e.g. a state) and we observe $\{Y(\mathbf{s}_i) : i = 1, \ldots, N\}$ at N distinct spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_N$.

Gaussian Process

Gaussian Process: A stochastic process where ANY finite collection of observed random variables follow a multivariate normal distribution.

More technical definition: For any set of $t_1, \ldots, t_N \in \mathcal{T}$, the vector $\mathbf{Y} = (Y(t_1), \ldots, Y(t_N))' \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y)$.

Defining a GP: To define Gaussian distribution, we only need to define what $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_X$ are.

Gaussian Process

The Mean of a GP: $\boldsymbol{\mu}$

1. Constant mean $\boldsymbol{\mu} = \mu \mathbf{1}_N$
2. If we also have covariates $\mathbf{X}(t_i) = (X_1(t_i), \ldots, X_P(t_i))'$ then we can say,

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

The Covariance of a GP: $\boldsymbol{\Sigma}_Y$

1. We need a correlation function, say $\rho(\cdot)$, so that observations $Y(t_i)$ and $Y(t_j)$ are strongly correlated when $|t_i - t_j|$ is small.
2. The correlation function, $\rho(\cdot)$, also needs to be "positive definite" so that $\boldsymbol{\Sigma}_Y$ is a valid covariance matrix.

Gaussian Process

Matern Correlation Function:

$$\begin{aligned}\text{Corr}(Y(t_i), Y(t_j)) &= \frac{1}{2^{\nu-1}\gamma(\nu)} \left(2\phi\sqrt{\nu}|t_i - t_j|\right)^{\nu} K_\nu(2\phi\sqrt{\nu}|t_i - t_j|) \\ &= \texttt{Matern}(|t_i - t_j|, \texttt{nu} = \nu, \texttt{alpha} = \phi)\end{aligned}$$

Properties:

1. ϕ : decay parameter. As ϕ increases, correlation (at a fixed distance) decreases.
2. ν : smoothness parameter. As ν increases, smoothness increases.
3. $\texttt{Matern}(0, \texttt{nu} = \nu, \texttt{alpha} = \phi) = 1$
4. Eff. Range : distance where correlation decays to 0.05.

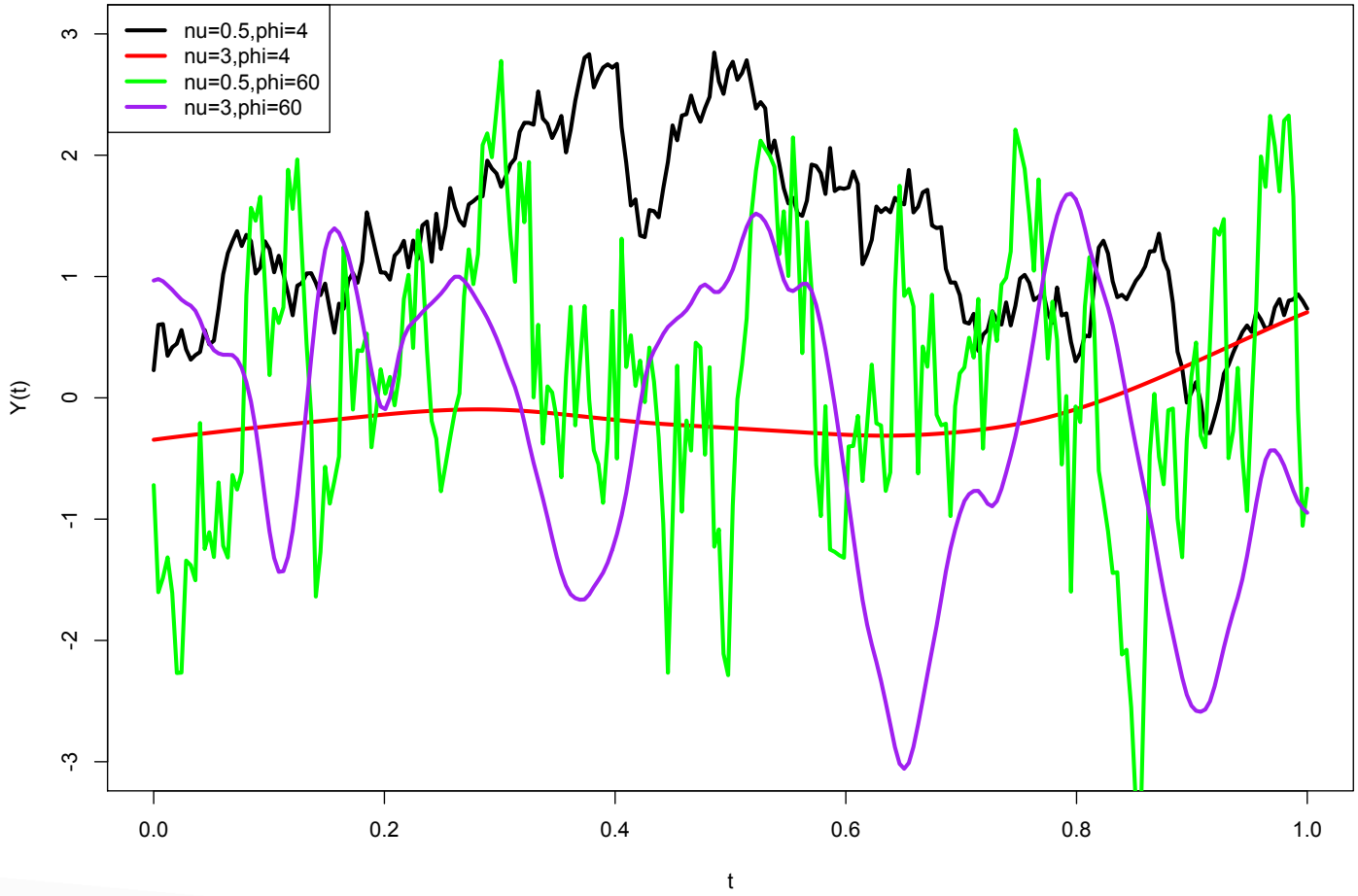
Gaussian Process

Matern Correlation Function:

$$\begin{aligned}\text{Corr}(Y(t_i), Y(t_j)) &= \frac{1}{2^{\nu-1}\gamma(\nu)} \left(2\phi\sqrt{\nu}|t_i - t_j|\right)^{\nu} K_{\nu}(2\phi\sqrt{\nu}|t_i - t_j|) \\ &= \texttt{Matern}(|t_i - t_j|, \texttt{nu} = \nu, \texttt{alpha} = \phi)\end{aligned}$$

Note: This is an example of an isotropic process (correlation only depends on Euclidean distance). There are other types but we won't cover those here.

Gaussian Process



Gaussian Process

Code for Simulating from a GP:

```
library(LatticeKrig) #Load rdist function
tseq <- seq(0,1,length=250) #Locations
D <- rdist(tseq) #Distance between locations
nu <- 0.5
phi <- 4
s2 <- 1
mu <- 0
V <- s2*Matern(D,alpha=phi,nu=nu) #Covariance matrix
draw <- mu+ t(chol(V))%*%rnorm(length(tseq)) #MVN draw
plot(tseq,draw,type="l",ylim=c(-3,3),lwd=3,xlab="t",ylab="Y(t)")
```

Gaussian Process Regression

Gaussian Process Regression: We observe $y(x_1), \ldots, y(x_N)$ at the "locations" $x_1, \ldots, x_N \in \mathcal{X}$. We want a function $w(x)$ so that:

1. $\sum_{i=1}^{N} (y(x_i) - w(x_i))^2$ is small.
2. $w(x)$ is smooth to prevent overfitting.

Big Idea: Let $w(x_1), \ldots, w(x_N)$ be random (model them) and use correlation to enforce smoothness.

Gaussian Process Regression

Big Idea in Mathematical Notation: Let,

$$w(x) \sim \mathcal{GP}(\mu, \rho(\cdot))$$

$$\Rightarrow \mathbf{W} = \begin{pmatrix} w(x_1) \\ \vdots \\ w(x_N) \end{pmatrix} \sim \mathcal{N}(\mu \mathbf{1}_N, \sigma^2 \mathbf{R})$$

where the ij^{th} element of $\mathbf{R}$ is,

$$\mathbf{R}_{ij} = \texttt{Matern}(|x_i - x_j|, \texttt{nu} = \nu, \texttt{alpha} = \phi)$$

Gaussian Process Regression

Full Gaussian Process Model: Let,

$$\mathbf{Y} \mid \mathbf{W} = \begin{pmatrix} y(x_1) \\ \vdots \\ y(x_N) \end{pmatrix} \sim \mathcal{N}\left(\mathbf{W}, \tau^2 \mathbf{I}_N\right)$$

$$\mathbf{W} = \begin{pmatrix} w(x_1) \\ \vdots \\ w(x_N) \end{pmatrix} \sim \mathcal{N}\left(\mu \mathbf{1}_N, \sigma^2 \mathbf{R}\right)$$

which marginalizes to:

$$\mathbf{Y} \sim \mathcal{N}\left(\mu \mathbf{1}_N, \sigma^2 \mathbf{R} + \tau^2 \mathbf{I}_N\right)$$

Gaussian Process Regression

Prediction for Gaussian Process Regression: Suppose we want to predict y at values $x_1^\star, \ldots, x_K^\star$. By the Gaussian process assumption,

$$\begin{pmatrix} \mathbf{Y}^\star \\ \mathbf{Y} \end{pmatrix} = \begin{pmatrix} y(x_1^\star) \\ \vdots \\ y(x_K^\star) \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}\left(\mu \mathbf{1}_N, \sigma^2 \begin{pmatrix} \mathbf{R}_{Y^\star} & \mathbf{R}_{Y^\star, Y} \\ \mathbf{R}'_{Y^\star, Y} & \mathbf{R}_Y \end{pmatrix} + \tau^2 \mathbf{I}\right)$$

So, the predictions for $\mathbf{Y}^\star$ are the expected values of the conditional distribution for $\mathbf{Y}^\star \mid \mathbf{Y}$,

$$\mathbb{E}(\mathbf{Y}^\star \mid \mathbf{Y}) = \mu \mathbf{1}_K + \sigma^2 \mathbf{R}_{Y^\star, Y} \left(\sigma^2 \mathbf{R}_Y + \tau^2 \mathbf{I}_N\right)^{-1} (\mathbf{Y} - \mu \mathbf{1}_N)$$

Gaussian Process Regression

Prediction Intervals using Gaussian Process Regression:

Due to the joint normal assumption, the 95% prediction interval is the 0.025 and 0.975 quantiles of the conditional distribution for $\mathbf{Y}^\star \mid \mathbf{Y}$,

$$\mathbf{Y}^\star \mid \mathbf{Y} \sim \mathcal{N}\left(\boldsymbol{\mu}_{Y^\star|Y}, \boldsymbol{\Sigma}_{Y^\star|Y}\right)$$

where,

$$\boldsymbol{\mu}_{Y^\star|Y} = \mu \mathbf{1}_K + \sigma^2 \mathbf{R}_{Y^\star,Y} \left(\sigma^2 \mathbf{R}_Y + \tau^2 \mathbf{I}_N\right)^{-1} (\mathbf{Y} - \mu \mathbf{1}_N)$$

$$\boldsymbol{\Sigma}_{Y^\star|Y} = (\sigma^2 \mathbf{R}_{Y^\star} + \tau^2 \mathbf{I}_K) - \left[\sigma^2 \mathbf{R}_{Y^\star,Y}\right] \left(\sigma^2 \mathbf{R}_Y + \tau^2 \mathbf{I}_K\right)^{-1} \left[\sigma^2 \mathbf{R}'_{Y^\star,Y}\right]$$

Gaussian Process Regression

Fitting a GP Model: We observe $y(x_1), \ldots, y(x_N)$, which leads to the likelihood,

$$\mathbf{Y} = \begin{pmatrix} y(x_1) \\ \vdots \\ y(x_n) \end{pmatrix} \sim \mathcal{N}\left(\mu \mathbf{1}_N, \sigma^2 \mathbf{R} + \tau^2 \mathbf{I}_N\right)$$

where the ij^{th} element of $\mathbf{R}$ is,

$$\mathbf{R}_{ij} = \texttt{Matern}(|x_i - x_j|, \texttt{nu} = \nu, \texttt{alpha} = \phi)$$

So, the unknown parameters are μ, σ^2, ν, ϕ and τ^2.

We estimate these by maximum likelihood.

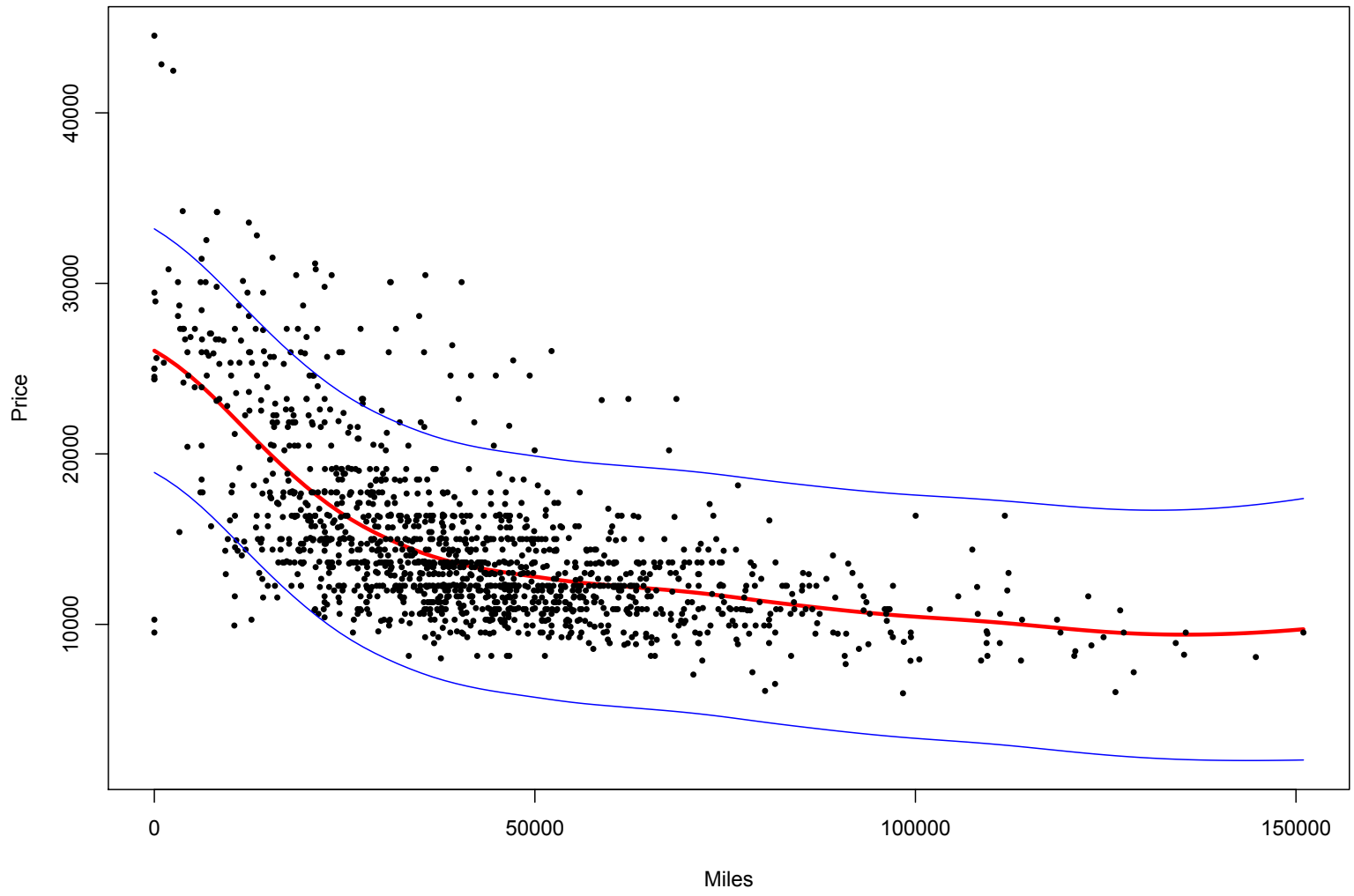
Gaussian Process Regression

Fitting a GP Model: Issues to be aware of:

1. The data often contains ZERO information about ν.
   - controls the number of derivatives so it is commonly set at 2 for a smooth function and 0.5 if you want a bumpy function (and stability).
2. The data often contains little information about ϕ.
   - Try to estimate it anyway.
   - Do maximum likelihood over a small grid of ranges.

Gaussian Process Regression in R

```
library(geoR)
library(LatticeKrig)
nu <- 2
N <- length(miles)
obs.data <- as.geodata(cbind(price,miles,rep(0,length(miles))),data.col=1,coords.col=2:3)
gp.fit <- likfit(obs.data,cov.model="matern",kappa=nu,fix.kappa=TRUE,
      ini.cov.pars=c(s2.start.val,phi.start.val), trend="cte")
phi <- 1/gp.fit$phi
s2 <- gp.fit$sigmasq
mu <- gp.fit$beta
tau2 <- gp.fit$tausq
K <- 100
pred.seq <- seq(min(miles),max(miles),length=K)
D <- rdist(c(pred.seq,miles))
V <- s2*Matern(D,alpha=phi,nu=nu) ##V = Sigma_Y
EV <- mu + V[1:K,K+(1:N)]%*%solve(V[K+(1:N),K+(1:N)]+tau2*diag(N))%*%(price-mu)
cond.Var <- diag((V[1:K,1:K]+tau2*diag(K))-V[1:K,K+(1:N)]%*%solve(V[K+(1:N),K+(1:N)]
+tau2*diag(N))%*%t(V[1:K,K+(1:N)]))
upper <- qnorm(0.975,mean=EV,sd=sqrt(cond.Var))
lower <- qnorm(0.025,mean=EV,sd=sqrt(cond.Var))
```




```
plot(pred.seq,EV,type="l",lwd=3,xlab="Miles",ylab="Price",xlim=range(miles),ylim=c(min(lower),ma
x(price)),col="red")
points(miles,price,pch=19,cex=0.5)
lines(pred.seq,lower,col="blue")
lines(pred.seq,upper,col="blue")
```

Gaussian Process Regression Multiple X's

GP Regression with Multiple X's:

1. GP Regression is used to deal with non-linearity.
   Most variables will just be part of the mean:

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

   or likfit(...,trend=~X) without a columns of 1's.

How would we do inference for $\boldsymbol{\beta}$?
This is just a generalized least-squares problem!

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$$

$$\mathbf{V} = \sigma^2\mathbf{R} + \tau^2\mathbf{I}_N$$

$$\Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

$$\text{SE}(\boldsymbol{\beta}) = \sqrt{(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}}$$

Gaussian Process Regression Multiple X's

GP Regression with Multiple X's:

2. When multiple X's have non-linearity, the model is

$$\mathbf{Y} = \begin{pmatrix} y(\mathbf{x}_1) \\ \vdots \\ y(\mathbf{x}_n) \end{pmatrix} \sim \mathcal{N}\left(\mu \mathbf{1}_N, \sigma^2 \mathbf{R} + \tau^2 \mathbf{I}_N\right)$$

where the ij^{th} element of $\mathbf{R}$ is,

$$\mathbf{R}_{ij} = \texttt{Matern}(\|\mathbf{x}_i - \mathbf{x}_j\|, \texttt{nu} = \nu, \texttt{alpha} = \phi)$$

or, better yet,

$$\mathbf{R}_{ij} = \prod_{p=1}^{P} \texttt{Matern}(|x_{ip} - x_{jp}|, \texttt{nu} = \nu_p, \texttt{alpha} = \phi_p)$$

Gaussian Processes in Spatial Statistics

Issue in Spatial Statistics: $y(\mathbf{s})$ is highly nonlinear as a function of spatial location $\mathbf{s}$.

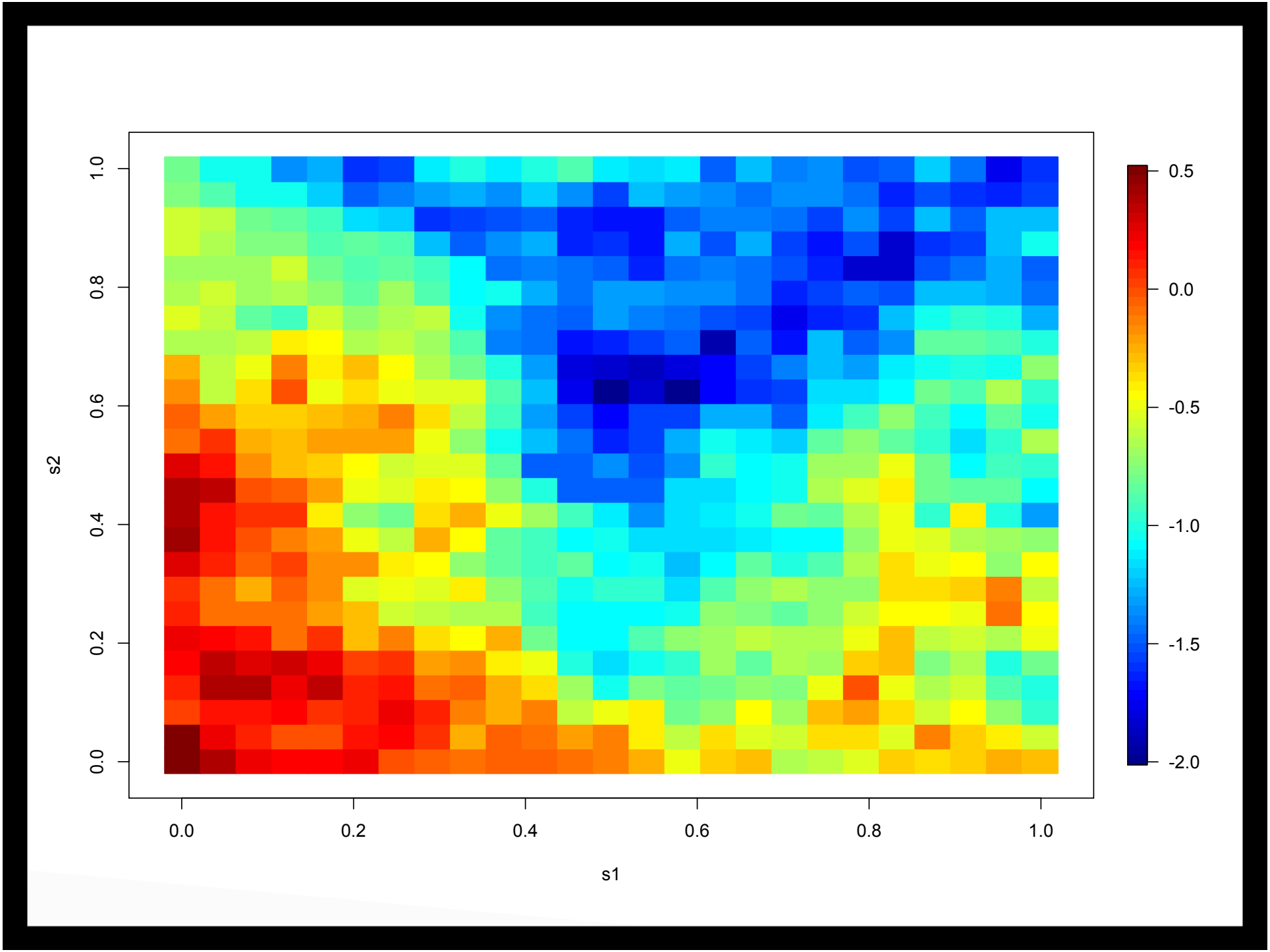
Spatial Statistics: We observe $y(\mathbf{s}_1), \ldots, y(\mathbf{s}_N)$ and the covariates $\mathbf{x}(\mathbf{s}_1), \ldots, \mathbf{x}(\mathbf{s}_N)$ at N distinct spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_N$ in some spatial region $\mathcal{D}$.

Spatial Statistics Model:

$$\mathbf{Y} = \begin{pmatrix} y(\mathbf{s}_1) \\ \vdots \\ y(\mathbf{s}_N) \end{pmatrix} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{R} + \tau^2\mathbf{I}_N\right)$$

where the ij^{th} element of $\mathbf{R}$ is,

$$\mathbf{R}_{ij} = \texttt{Matern}(\|\mathbf{s}_i - \mathbf{s}_j\|, \texttt{nu} = \nu, \texttt{alpha} = \phi)$$


s2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
s1
0.5
0.0
-0.5
-1.0
-1.5
-2.0

Gaussian Process in Spatial Statistics

```
g.size <- 25
s <- as.matrix(expand.grid(seq(0.001,0.999,length=g.size),seq(0.001,0.999,length=g.size)))
mu <- 0
s2 <- 1^2
tau2 <- 0
phi <- 0.5
nu <- 0.5
D <- rdist(s)
V <- tau2*diag(nrow(s))+s2*Matern(D,alpha=phi,nu=nu)
y <- mu+t(chol(V))%*%rnorm(nrow(V))
image.plot(matrix(s[,1],nrow=g.size), matrix(s[,2],nrow=g.size),
      matrix(y,nrow=g.size), xlab="s1",ylab="s2")
```

Gaussian Processes in Spatial Statistics

Spatial Kriging (Prediction): Calculate conditional expectation at new locations given observations.

Spatial Uncertainty Quantification: Variance at new location conditional on observations.

(see Slides 16-17)

Good and Bad of GPs

The Good:

1. Extremely flexible – can fit a wide variety of nonlinear functions.
2. Have predictable (AWESOME) tail behavior – process just reverts to the mean outside the range of the data.
3. Fits the function to all the data - we don't have stupid knots so everything can be easily estimated using maximum likelihood.
4. We can incorporate other predictors (covariates) using $\mathbf{X}\beta$.

Good and Bad of GPs

The Bad:

1. Curse of Dimensionality can kill em – the GP relies on correlating things that are "close" but, in high dimensions, data are spread out and nothing is "close."
2. Computational nightmare to fit for large data sets – inverting a large matrix (this is an open research question).

Expectations for Agriculture Case Study

1. Give an overview of the problem.
2. Give a "bird's eye" view of GP regression and how it is going to help with solve the problem.
3. Write out a GP model for the agriculture data.
   - Be sure to explicitly state what the correlation between observations are (e.g. Matern)
   - Be sure to define/explain any parameters that you use.
4. Fit your GP model and report parameter estimates.
   - Point estimates of covariance function parameters are fine but see if you can get a CI for the mean.
5. Show a fitted curve on how CWSI relates to SWC.
6. Give some measure of predictive accuracy (coverage and width). Note, you don't have a ton of data so maybe try leave-one-out cross validation.