Applied Spatial Statistics: Lecture 8
Conditional Inference Cross Validation

Douglas Nychka,
National Center for Atmospheric Research

Supported by the National Science Foundation *Boulder, Spring 2013*

Outline

- Standard errors
- Generating a conditional sample
- Simultaneous confidence bands
- CO example
- Cross-validation

Inference: *a conclusion reached on the basis of evidence and reasoning.*

Estimating a curve or surface.

The additive statistical model:

Given n pairs of observations (x_i, y_i), $i = 1, \ldots, n$

$$y_i = f(x_i) + \epsilon_i$$

ϵ_i's are random errors – usually $\boldsymbol{e} \sim (0, \sigma^2 I)$

$$f(\boldsymbol{x}) = \sum_{j=1}^{p} \phi_j(\boldsymbol{x})\boldsymbol{d}_j + g(\boldsymbol{x})$$

- $\{\phi_j\}$ *known* basis functions based on spatial locations
e.g. polyonimials or topography. $\boldsymbol{d}$ fixed coefficients
- $g(\boldsymbol{x})$ a Gaussian process expected value of 0 and a covariance function
$\rho k_\theta(\boldsymbol{x}_1, \boldsymbol{x}_2)$

- As a linear model: $\boldsymbol{y} = X\boldsymbol{d} + (\boldsymbol{g} + \boldsymbol{e})$

Distribution of prediction error

Prediction variance

$$\left(\hat{f}(\boldsymbol{x}^*) - f(\boldsymbol{x}^*)\right) \sim N(0, v)$$

v : a complicated function of the covariance for g, σ^2 and X.

- Can be derived using the *linear statistics rules*.
- Computing value of v is as much work as a full spatial prediction

Predictions at many locations:

$$\begin{pmatrix} \hat{f}(\boldsymbol{x}_1^*) - f(\boldsymbol{x}_1^*) \\ \hat{f}(\boldsymbol{x}_2^*) - f(\boldsymbol{x}_2^*) \\ \vdots \\ \hat{f}(\boldsymbol{x}_m^*) - f(\boldsymbol{x}_m^*) \end{pmatrix} \sim N(0, \boldsymbol{V})$$

- $\{\boldsymbol{x}_k^*\}$ possibly a huge grid of m points
- $\boldsymbol{V}$: an $m \times m$ covariance *matrix*.

How to take advantage of V?

Probabilities for the bivariate normal

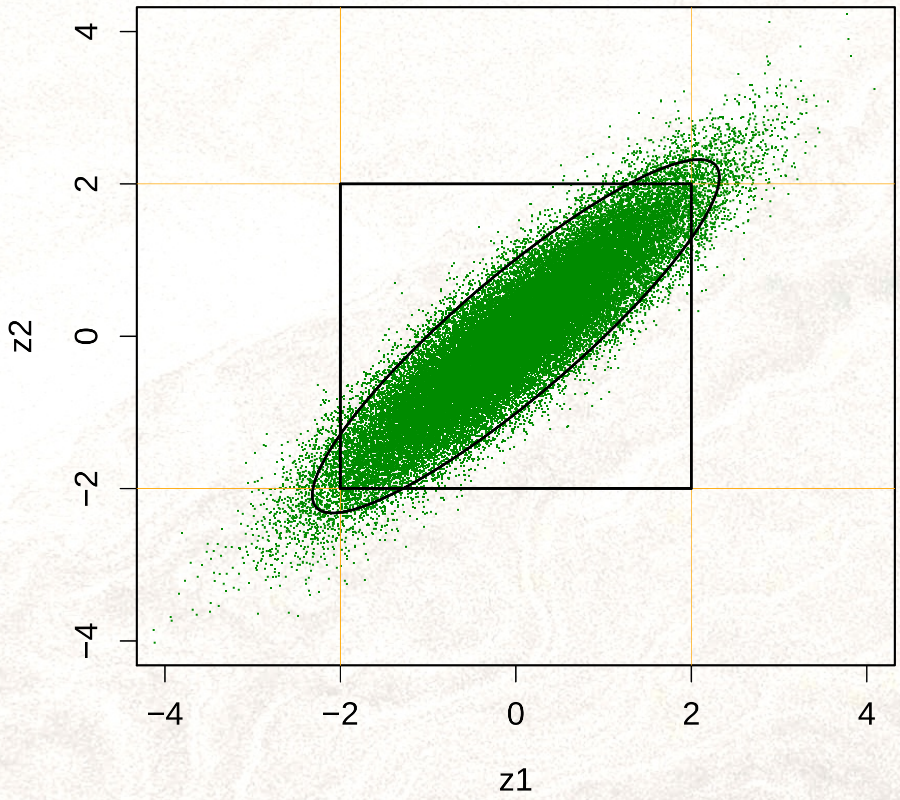
Bivariate normal:

$\boldsymbol{z}$ is multivariate Gaussian with mean vector 0, variances 1 and correlation .8.

Mean vector and Covariance matrix:

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix} \tag{1}$$

Draw of 50000 samples



Both sets contain about 93% percent of points. Elliptical contours are more efficient but also more complicated in shape.

Sampling from error distribution

Main idea:
Draw a sample from $\boldsymbol{e}^* \sim N(0, \boldsymbol{V})$

- $\hat{\boldsymbol{f}} + \boldsymbol{e}^*$ is interpreted as a " likely " value for $\boldsymbol{f}$ given the observations.

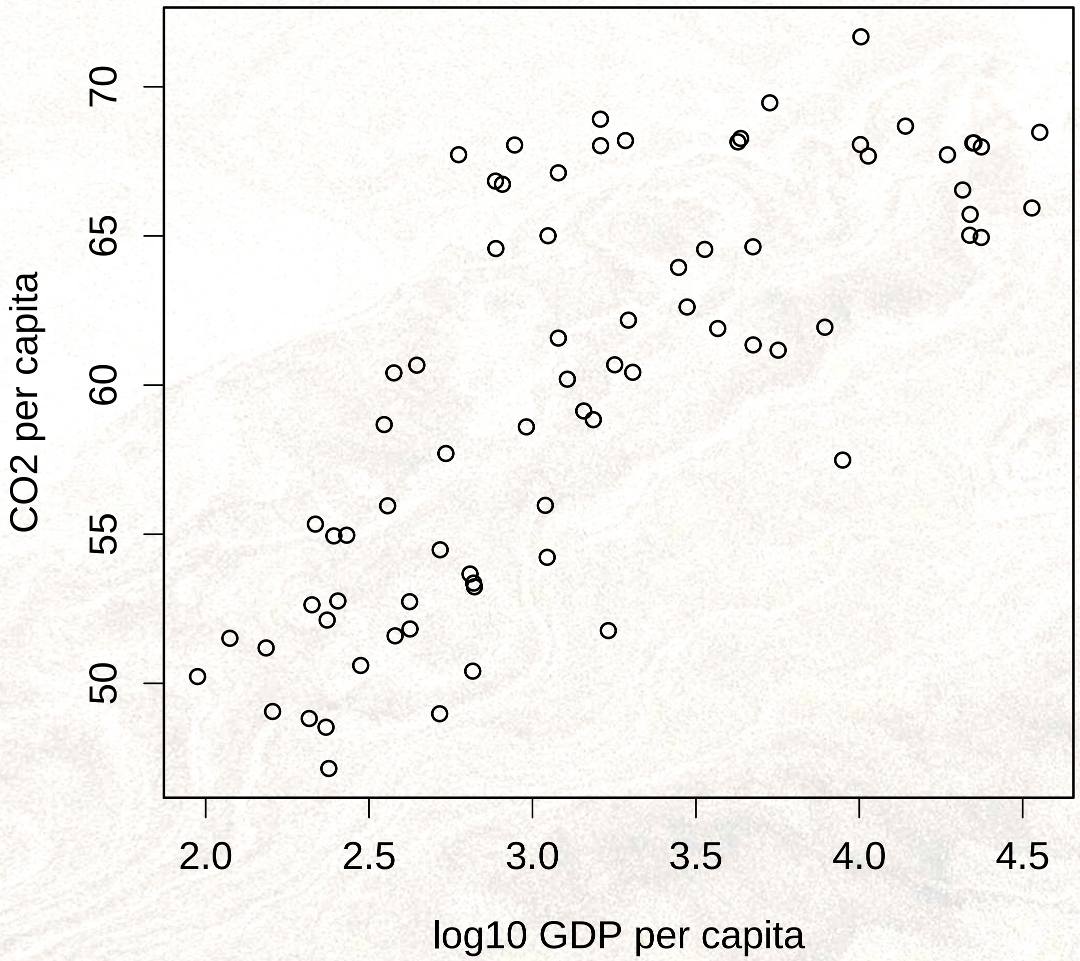
- Use cholesky method to generate random sample from errors

- Can apply nonlinear functions to a sample of $\hat{\boldsymbol{f}} + \boldsymbol{e}^*$ to make complicated inferences.

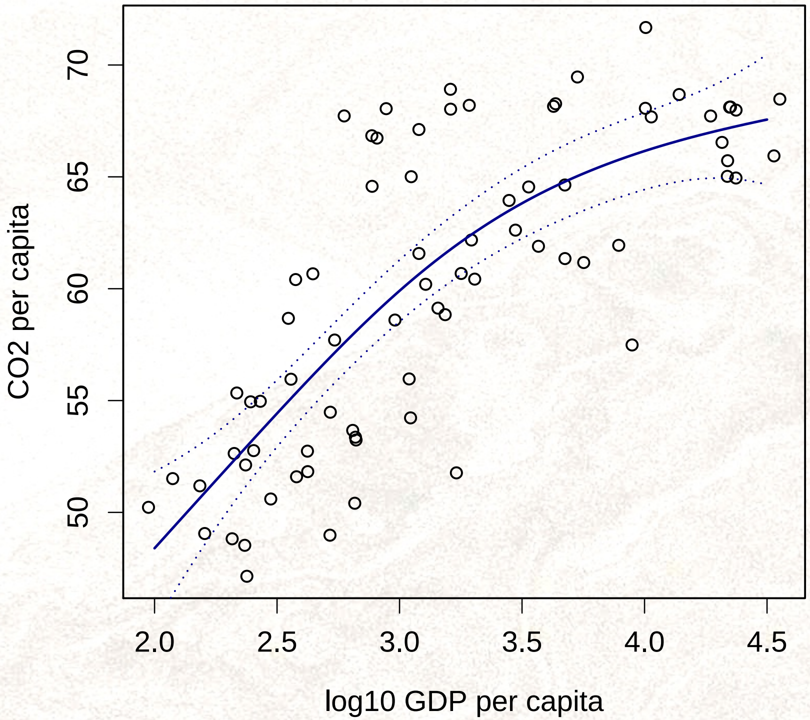
e.g. A confidence interval for the *maximum* value of the function or its derivative.

e.g. Uncertainty range for a contour of the predicted surface.

World Bank CO_2 data



Fit using `Krig`



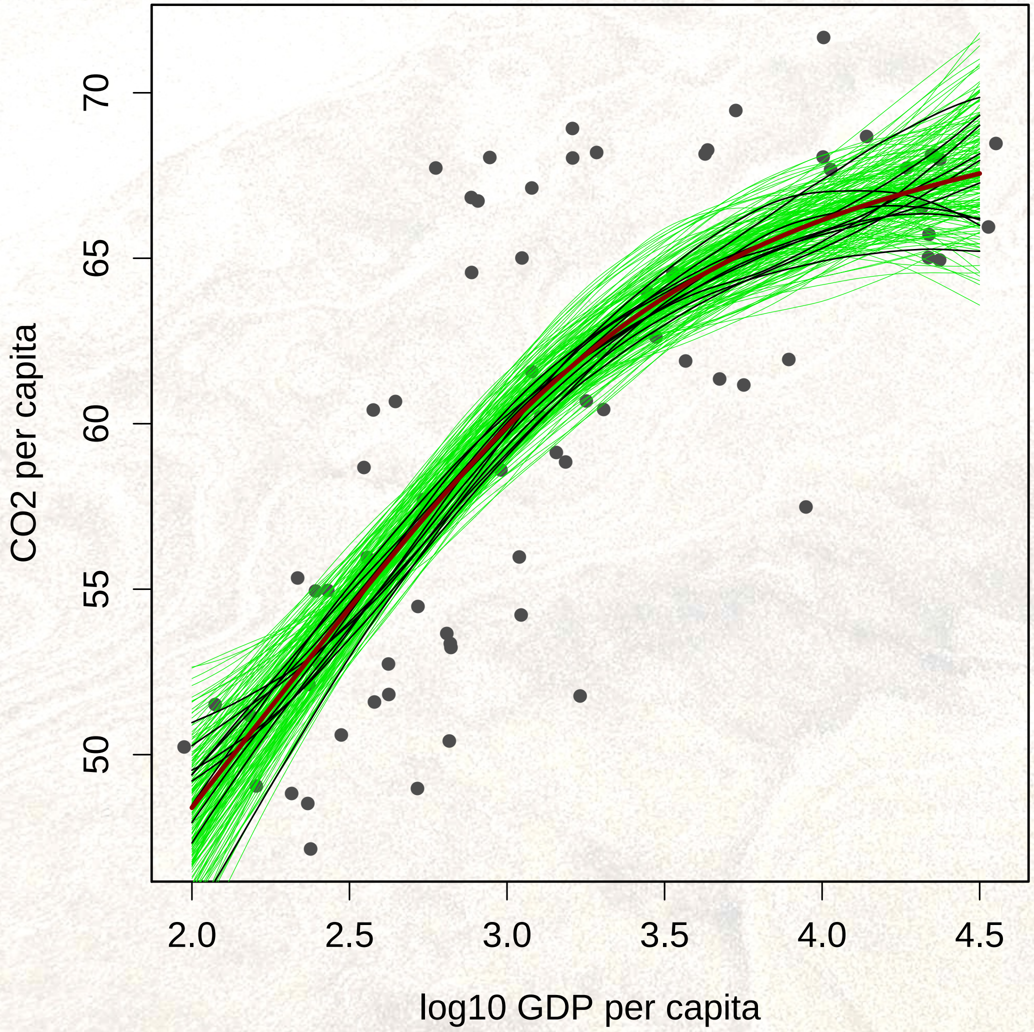

```
  fit.CO2<-
     Krig( x,y, Covariance="Matern",
             smoothness =2.0, theta=4.0,
             method="REML")
# fine grid of points
  xg<- seq( 2, 4.5,,150)
  fhat<- predict( fit.CO2, xg)
  segrid<- predict.se( fit.CO2, xg)
```

Each vertical range should be interpreted as a 95% confidence interval for f.

This is *not* a 95% confidence band

Sampling prediction errors

```
bigV<- predict.se( fit.CO2, xg, cov=TRUE)
chol( bigV)-> V.chol
bigE<- t( V.chol)%*%matrix( rnorm(200*150), ncol=200, nrow=150)
fstar<-  c(fhat)  + E
```



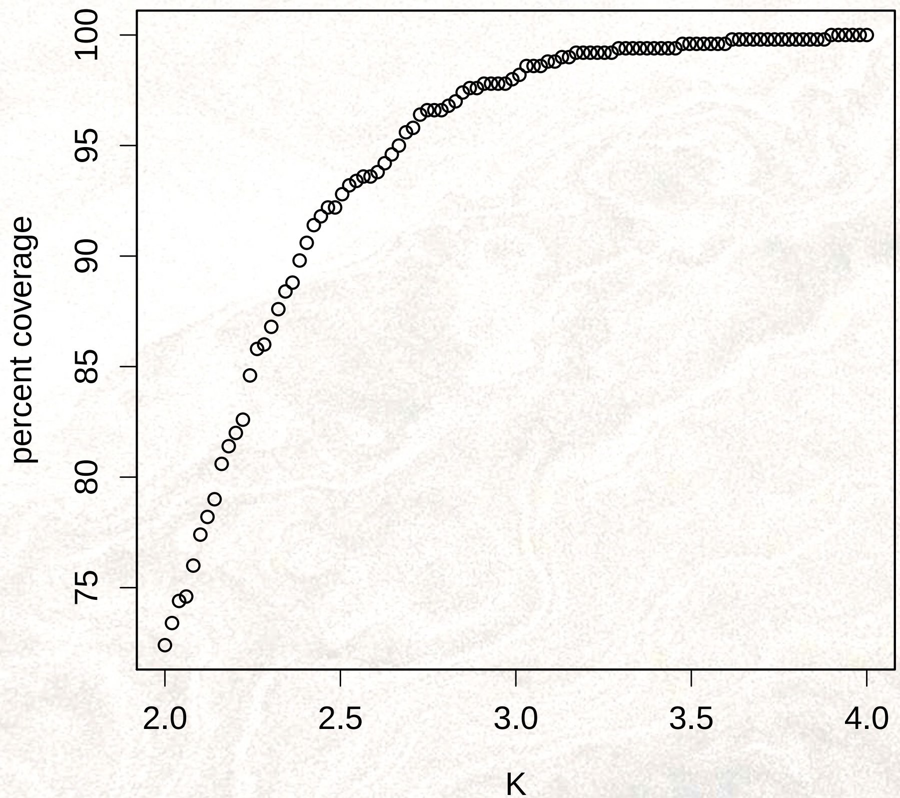
Computing simultaneous bands

Inflate the pointwise intervals to give simultaneous coverage

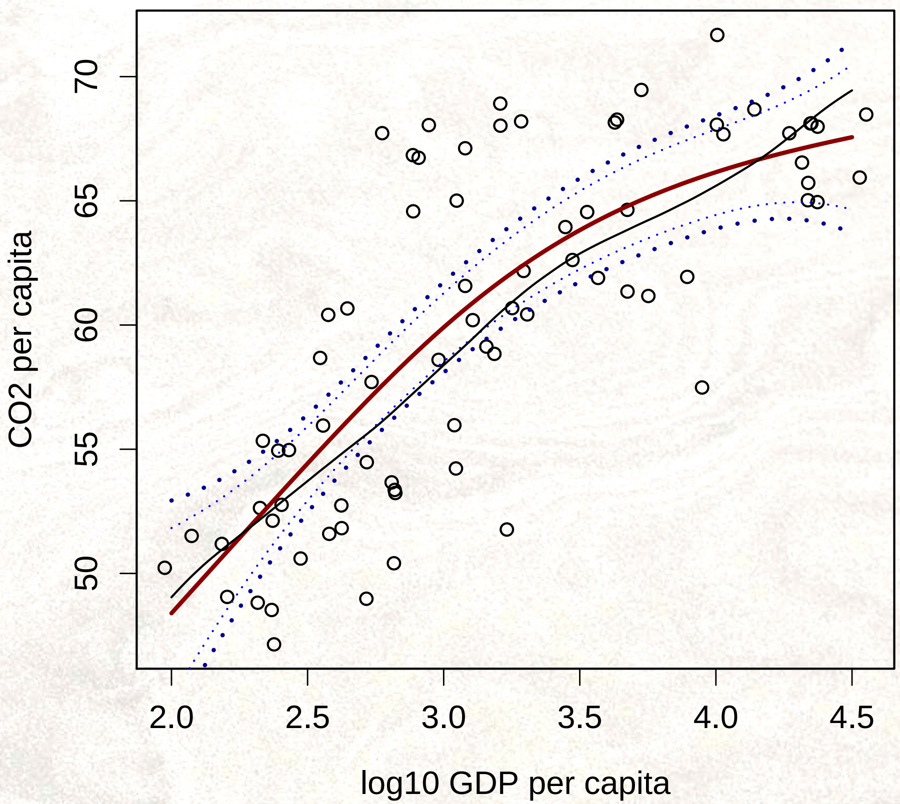
$$\mathcal{I}(x) = \hat{f}(x) \pm K * SE(x)$$

Determine K so that probability that $f(x) \in \mathcal{I}(x)$ for all x is 95%. K can can be estimated from conditional sample (500 members in example below).

Finding K

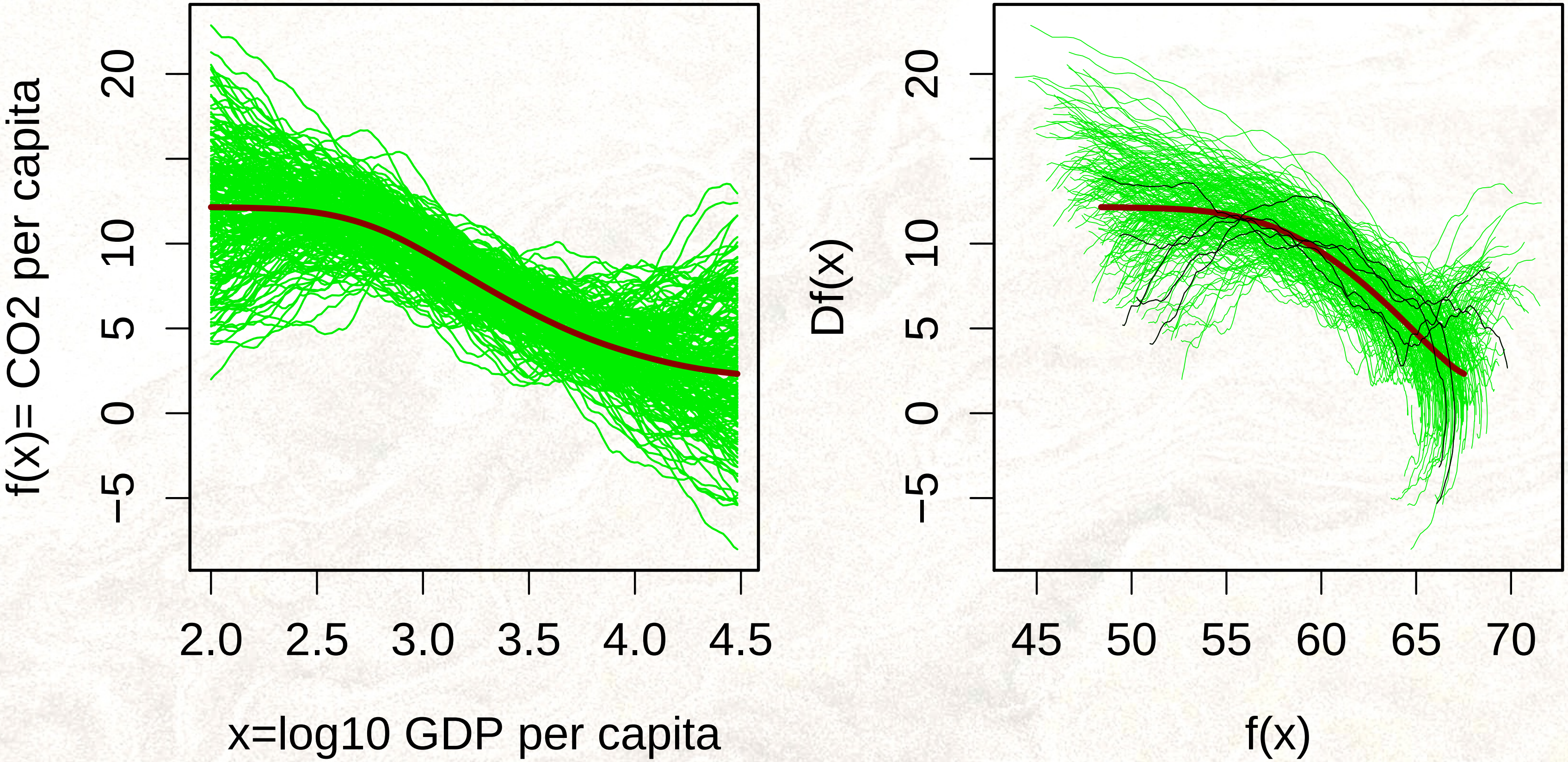


95% Bands $K = 2.65$

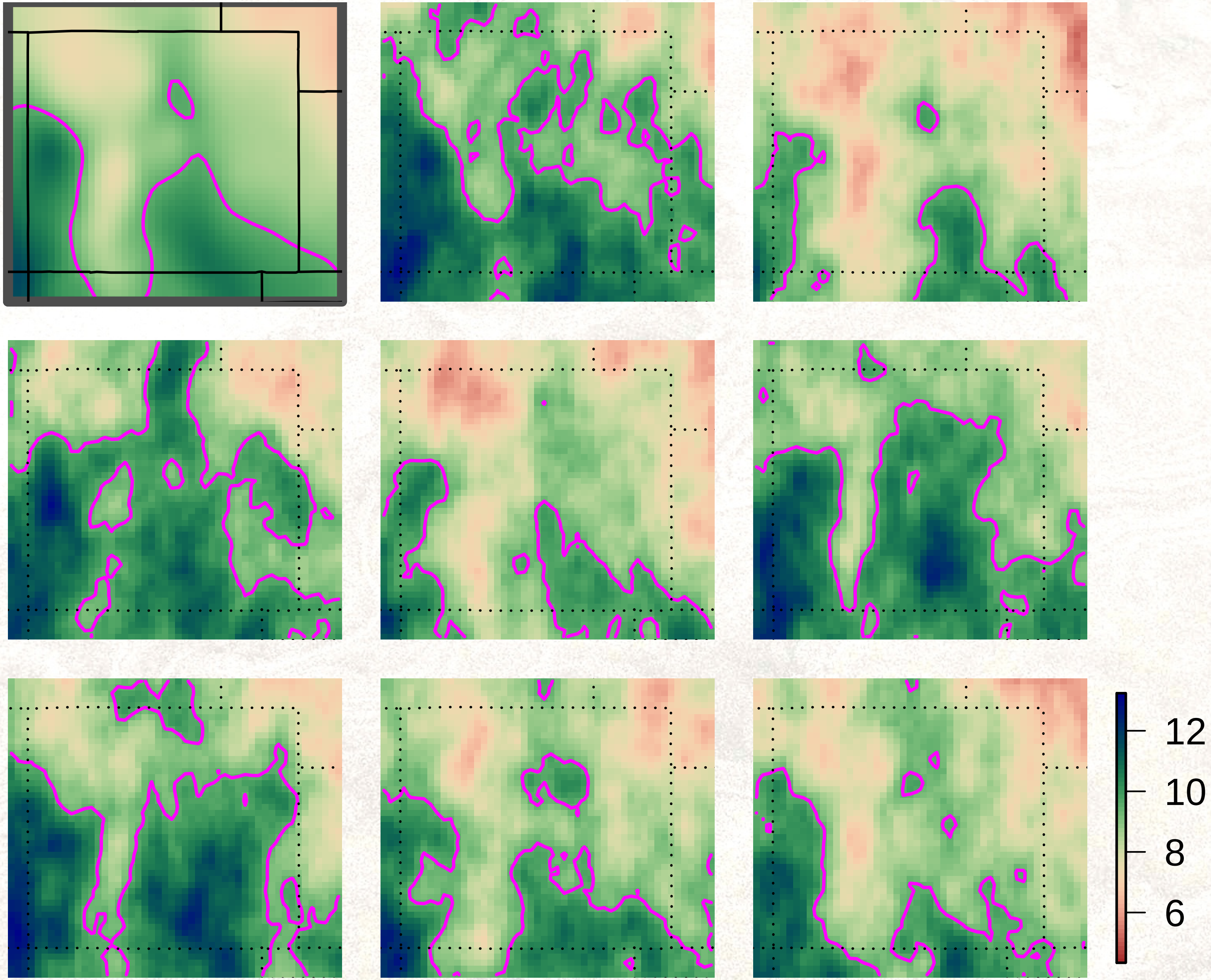


Looking at the derivative

$f'(x)$ and estimating the function: $f' = \Phi(f)$



Ensemble for Colorado MAM temps



Cross-validation

Omit subsets of data and compare them to the predictions using the remaining dataset.

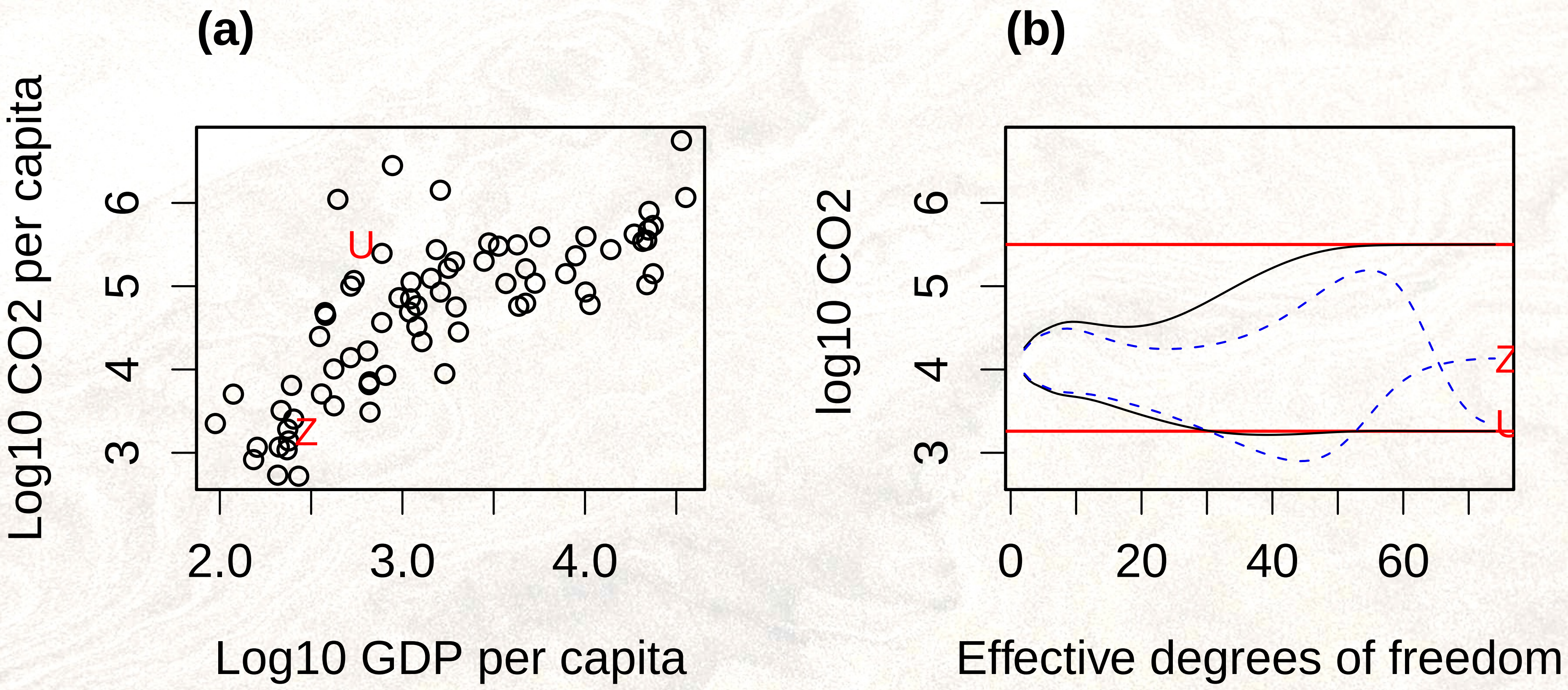
- Covariance model can be fixed
- Parameters can be varied for each subset/case

Leave-one-out cross-validation

Sequentially leave each data value out and find its prediction. The prediction in this case is "out of sample" because data point is not used.

World Bank example

Varying the smoothing parameter λ and comparing to omitted data points



$\widehat{f}^i$: spatial prediction omitting the ith data point

Cross-validation residual: $R_i = \boldsymbol{y}_i - \widehat{f}^i(\boldsymbol{x}_i)$

Diagnostic for inference: $R_i/\sqrt{SE(\boldsymbol{x}_i)^2 + \sigma^2)}$ should be $N(0,1)$