# Statistical Modeling SDS 383D: Excercise 6

Khai Nguyen - kbn397
April 24, 2022

**Curve fitting by linear smoothing** :
(A) We have the LSE estimate $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$. Therefore

$$\begin{aligned}\hat{\beta} x^* &= \left(X^T X\right)^{-1} X^T Y x^* \\ &= \left(\sum_{i=1}^{n} x_i^T x_i\right)^{-1} \sum_{i=1}^{n} x_i^T y_i x^*\end{aligned}$$

We can see that

$$w\left(x_i, x^*\right) = \frac{x_i^T}{x_i^T x_i} x^*.$$

The above equation smooths the new $x^*$ by scaling it with the ratio $\frac{x_i^T}{x_i^T x_i}$. This effectively take the proximity of $x^*$ to each $x_i$ into account when summing over all $x_i$. In comparison, the K-nearest-neighbor smoothing simply scales $x^*$ uniformly.
(B) I choose function $f(x) = xcos(x)$ and $\epsilon \sim \mathcal{N}(0, 4)$. I plot estimated functions with bandwidth $h \in \{0.1, 1, 2, 5, 10\}$ in Figure 1.

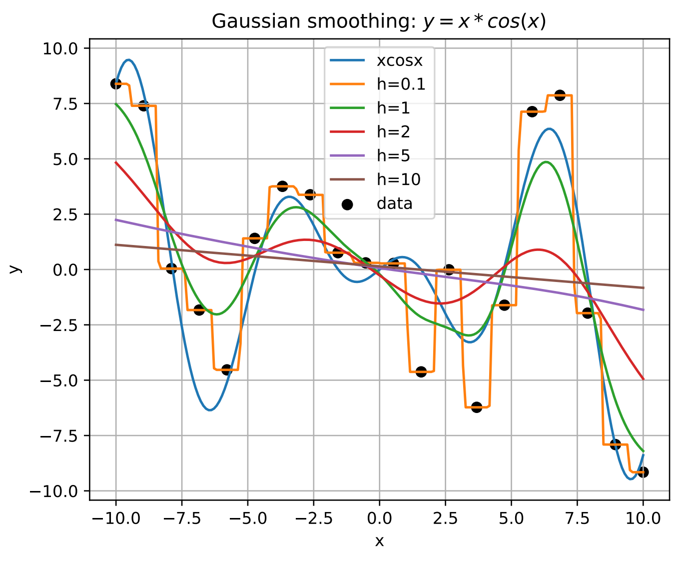



Figure 1: Estimated functions with various bandwidths.

**Cross Validation** :

(A) For the setting in the previous section, I got the MSE of $[5.07, 6.04, 13.93, 19.76, 20.47]$ for $h \in \{0.1, 1, 2, 5, 10\}$.

(B) I choose the following settings:

| | High Noise | Low Noise |
|---|---|---|
| Wiggly Function | $f(x) = x\cos(x) + \mathcal{N}(0, 10)$ | $f(x) = x\cos(x) + \mathcal{N}(0, 2)$ |
| Smooth Function | $f(x) = x^3 + x^2 + x + 1 + \mathcal{N}(0, 200)$ | $f(x) = x^3 + x^2 + x + 1 + \mathcal{N}(0, 2)$ |

I set the bandwidth $h \in \{0.1, 1, 2, 5, 10\}$, I got the following values

| | Wiggly/High | Wiggly/Low | Smooth/High | Smooth/Low |
|---|---|---|---|---|
| $h = 0.1$ | 18.43 | 1.01 | 4483.37 | 5.09 |
| $h = 1$ | 4.99 | 2.27 | 1683.32 | 855.61 |
| $h = 2$ | 12.69 | 10.47 | 9323.79 | 7915.5 |
| $h = 5$ | 17.25 | 16.20 | 40900.89 | 39829.34 |
| $h = 10$ | 17.61 | 17.10 | 85595.63 | 84566.26 |

We can see that when the noise level is low, a smaller bandwidth $h$ gives a better MSE. In contrast, the previous statement is not true when the noise level is high. The out-of-sample predictive validation seems to lead to reasonable choice of $h$ (see Figure 2).

(D) I use the same setting as previous parts. Using leave-one-out lemma, I get the following table:

| | Wiggly/High | Wiggly/Low | Smooth/High | Smooth/Low |
|---|---|---|---|---|
| $h = 0.1$ | 109.42 | 4.50 | 43052.47 | 9.11 |
| $h = 1$ | 101.35 | 6.60 | 41851.16 | 2026.89 |
| $h = 2$ | 111.68 | 15.28 | 49710.43 | 10290.67 |
| $h = 5$ | 119.12 | 22.01 | 81728.89 | 44710.91 |
| $h = 10$ | 119.71 | 22.78 | 129908.67 | 91568.80 |

The table gives a consistent result for best choice of $h$ to the out-of-sample predictive validation.

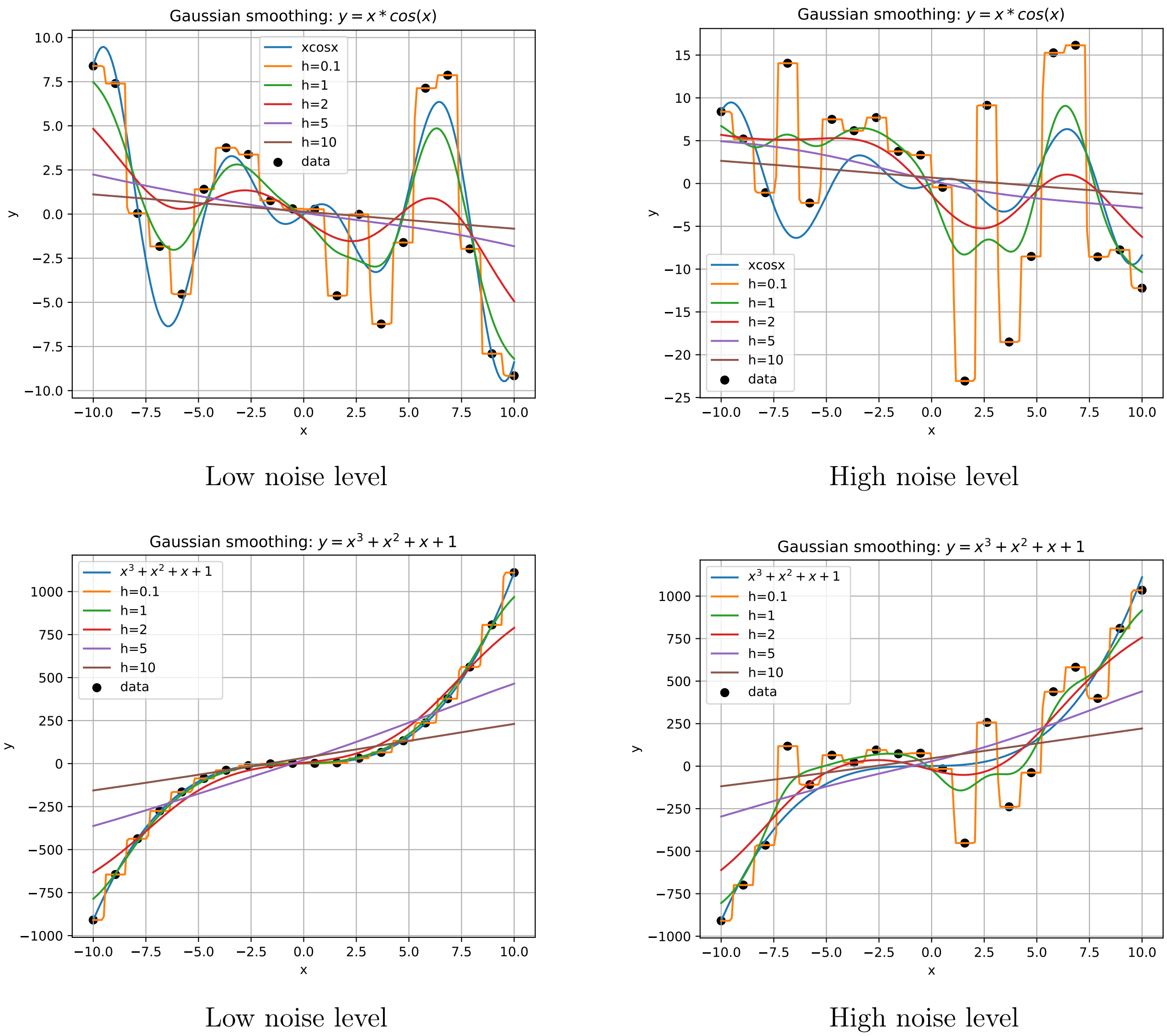
**Local Polynomial Regression** :

(A) We have the matrix form

$$\begin{aligned} R_x \mathbf{a} &= \begin{bmatrix} a_0 + a_1(x_1 - x) + \ldots + a_D(x_1 - x)^D \\ \vdots \\ a_0 + a_1(x_n - x) + \ldots + a_D(x_n - x)^D \end{bmatrix} \\ &= \begin{bmatrix} g_x(x_1 \mid \mathbf{a}) \\ \vdots \\ g_x(x_n \mid \mathbf{a}) \end{bmatrix} \end{aligned}$$

Therefore

$$\sum_{i=1}^{n} \tilde{w}_i \{y_i - g_x(x_i; a)\}^2 = (y - R_x \mathbf{a})^T \operatorname{diag}(\tilde{\mathbf{w}})(y - R_x \mathbf{a})$$




Low noise level

High noise level

Low noise level

High noise level

Figure 2: Estimated functions with various bandwidths and noise levels.

Taking the derivative

$$\frac{\partial}{\partial \mathbf{a}} \left[ (y - R_x \mathbf{a})^T \operatorname{diag}(\tilde{\mathbf{w}}) (y - R_x \mathbf{a}) \right] = -\left( y^T \operatorname{diag}(\tilde{\mathbf{w}}) R_x \right)^T - R_x^T \operatorname{diag}(\tilde{\mathbf{w}}) y + 2 \left( R_x^T \operatorname{diag}(\tilde{\mathbf{w}}) R_x \right) \mathbf{a}$$
$$= -2 R_x^T \operatorname{diag}(\tilde{\mathbf{w}}) y + 2 \left( R_x^T \operatorname{diag}(\tilde{\mathbf{w}}) R_x \right) \mathbf{a}$$

Setting the derivative to 0, we have

$$R_x^T \operatorname{diag}(\tilde{\mathbf{w}}) R_x \mathbf{a} = R_x^T \operatorname{diag}(\tilde{\mathbf{w}}) y$$
$$\hat{\mathbf{a}} = \left( R_x^T \operatorname{diag}(\tilde{\mathbf{w}}) R_x \right)^{-1} R_x^T \operatorname{diag}(\tilde{\mathbf{w}}) y$$

We have $\hat{f}(x) = \mathbf{e}^T \hat{\mathbf{a}}$.
(B) With $D = 1$, $g_x(x_i \mid a) = a_0 + a_1(x_i - x)$, we have

$$
\begin{aligned}
R_x^T \operatorname{diag}(\tilde{\mathbf{w}}) &= \begin{bmatrix} 1 & \dots & 1 \\ x_1 - x & \dots & x_n - x \end{bmatrix} \begin{bmatrix} \tilde{w}_1 & \dots & 0 \\ \vdots & \tilde{w}_i & \vdots \\ 0 & \dots & \tilde{w}_n \end{bmatrix} \\
&= \begin{bmatrix} \tilde{w}_1 & \dots & \tilde{w}_n \\ \tilde{w}_1(x_1 - x) & \dots & \tilde{w}_n(x_n - x) \end{bmatrix}, \\
R_x^T \operatorname{diag}(\tilde{\mathbf{w}}) R_x &= \begin{bmatrix} \tilde{w}_1 & \dots & \tilde{w}_n \\ \tilde{w}_1(x_1 - x) & \dots & \tilde{w}_n(x_n - x) \end{bmatrix} \begin{bmatrix} 1 & x_1 - x \\ \vdots & \vdots \\ 1 & x_n - x \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^n \tilde{w}_i & \sum_{i=1}^n \tilde{w}_i(x_i - x) \\ \sum_{i=1}^n \tilde{w}_i(x_i - x) & \sum_{i=1}^n \tilde{w}_i(x_i - x)^2 \end{bmatrix}, \\
\left(R_x^T \operatorname{diag}(\tilde{\mathbf{w}}) R_x\right)^{-1} &= \frac{1}{\mathcal{D}} \begin{bmatrix} \sum_{i=1}^n \tilde{w}_i(x_i - x)^2 & -\sum_{i=1}^n \tilde{w}_i(x_i - x) \\ -\sum_{i=1}^n \tilde{w}_i(x_i - x) & \sum_{i=1}^n \tilde{w}_i \end{bmatrix},
\end{aligned}
$$

where

$$
\begin{aligned}
\mathcal{D} &= \sum_{i=1}^n \tilde{w}_i(x_i - x)^2 - \left(\sum_{i=1}^n \tilde{w}_i(x_i - x)\right)^2 \\
&= \sum_{i=1}^n K(\cdot)(x_i - x)^2 - \left(\sum_{i=1}^n K(\cdot)(x_i - x)\right)^2 \\
&= s_2(x) - s_1^2(x)
\end{aligned}
$$

Let $S_k^{-1} = 1/\sum_{i=1}^n K(\cdot)$, we have

$$
\begin{aligned}
&\begin{bmatrix} 1 & 0 \end{bmatrix} \frac{1}{\mathcal{D}} \begin{bmatrix} \sum_{i=1}^n \tilde{w}_i(x_i - x)^2 & -\sum_{i=1}^n \tilde{w}_i(x_i - x) \\ -\sum_{i=1}^n \tilde{w}_i(x_i - x) & \sum_{i=1}^n \tilde{w}_i \end{bmatrix} \\
&\quad = \begin{bmatrix} \dfrac{\sum_{i=1}^n K(\cdot)(x_i - x)^2}{\mathcal{D}} & \dfrac{-\sum_{i=1}^n K(\cdot)(x_i - x)}{\mathcal{D}} \end{bmatrix}
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\hat{f}(x) &= \begin{bmatrix} \dfrac{\sum_{i=1}^n K(\cdot)(x_i - x)^2}{\mathcal{D}} & \dfrac{-\sum_{i=1}^n K(\cdot)(x_i - x)}{\mathcal{D}} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n K(\cdot) y_i \\ \sum_{i=1}^n K(\cdot)(x_i - x) y_i \end{bmatrix} \\
&= \frac{s_2(x) \sum_{i=1}^n K(\cdot) y_i - s_1(x) \sum_{i=1}^n K(\cdot)(x_i - x) y_i}{s_2(x) - s_1^2(x)} \\
&= \frac{\sum_{i=1}^n K(\cdot)\left[s_2(x) - (x_i - x) s_1(x)\right] y_i}{s_2(x) \sum_{i=1}^n K(\cdot) - s_1(x) \sum_{i=1}^n K(\cdot)(x_i - x)} \\
&= \frac{\sum_{i=1}^n w_i(x) y_i}{\sum_{i=1}^n w_i(x)}
\end{aligned}
$$

(C) We have

$$\begin{aligned}\hat{f}(x_i) &= e^T \hat{a} \\ &= e^T \left(X^T W X\right)^{-1} X^T W Y\end{aligned}$$

Therefore

$$\begin{aligned}E[\hat{f}(x)] &= e^T \left(X^T W X\right)^{-1} X^T W f(x), \\ \text{var}[\hat{f}(x)] &= \sigma^2 e^T \left(X^T W X\right)^{-1} X^T W W^T X \left(X^T W X\right)^{-1} e\end{aligned}$$

Simplify model in (B)

$$\begin{aligned}E[\hat{f}(x)] &= \tilde{w}^T f(x) \\ \text{var}[\hat{f}(x)] &= \sigma^2 \tilde{w}^T \tilde{w}\end{aligned}$$

where $\tilde{w} = \left[\frac{w_1(x)}{\sum_i w_i(x)}, \ldots, \frac{w_n(x)}{\sum_i w_i(x)}\right]^T$.
(D) By the given definition, we

$$\begin{aligned}\hat{\sigma}^2 &= \frac{(y - Hy)^T (y - Hy)}{n - 2\,\text{tr}(H) + \text{tr}\left(H^T H\right)} \\ &= \frac{(y - Hy)^T (y - Hy)}{\text{tr}\left[(I - H)^T (I - H)\right]}\end{aligned}$$

Hence

$$\begin{aligned}E[\hat{\sigma}^2] &= \frac{E[(y - Hy)^T (y - Hy)]}{\text{tr}\left[(I - H)^T (I - H)\right]} \\ &= \frac{\text{tr}\left((I - H)^T (I - H)\sigma^2\right) + \mu^T (I - H)^T (I - H)\mu}{\text{tr}\left[(I - H)^T (I - H)\right]} \\ &= \frac{\text{tr}\left((I - H)^T (I - H)\sigma^2\right) + \|f(x) - Hf(x)\|_2^2}{\text{tr}\left[(I - H)^T (I - H)\right]}\end{aligned}$$

which is unbiased for $\sigma^2$ when $\|f(x) - \text{Hf}(x)\|_2^2 = 0$.
(E) I use leave-out-out lemma to find the best value of $h$ among 100 values of $h \in [0.1, 100]$. I got $h^\star = 6.9$. I plot the estimated function in Figure 3.
(F) The residuals from the fitted model can be seen in Figure 4. From the figures, we can see that heteroskedasticity might be a better assumption here.
(G) I show the additional confident interval $\hat{f}(x) \pm 2 \cdot \sqrt{\hat{\sigma}^2 \|h\|_2^2}$ in Figure 5.

**Gaussian Processes** :
(A) I plot 10 random samples from GP with two covariance functions (different settings of $\tau_1, \tau_2, b$) in Figure 6-Figure 8. In particular, I change value of $b$ in Figure 6, value of $\tau_1$ in Figure 7, and value of $\tau_2$ in Figure 8. From these figures, $b$ acts like the bandwidth, $\tau_1$ controls the scales, and $\tau_2$ also

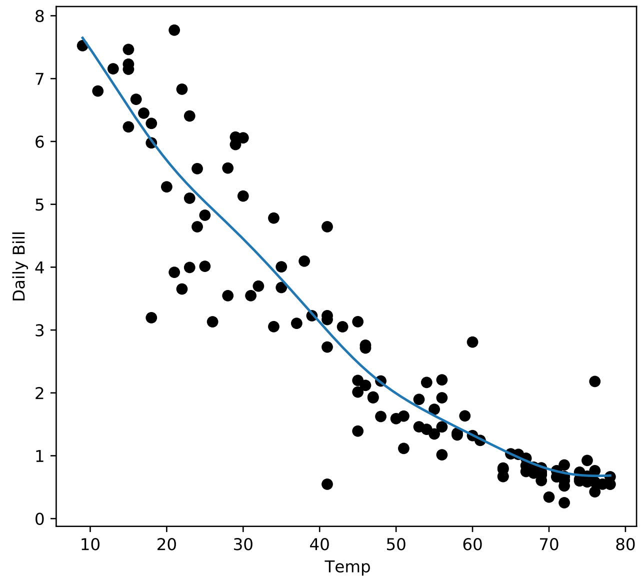



Figure 3: Estimated functions with local polynomial

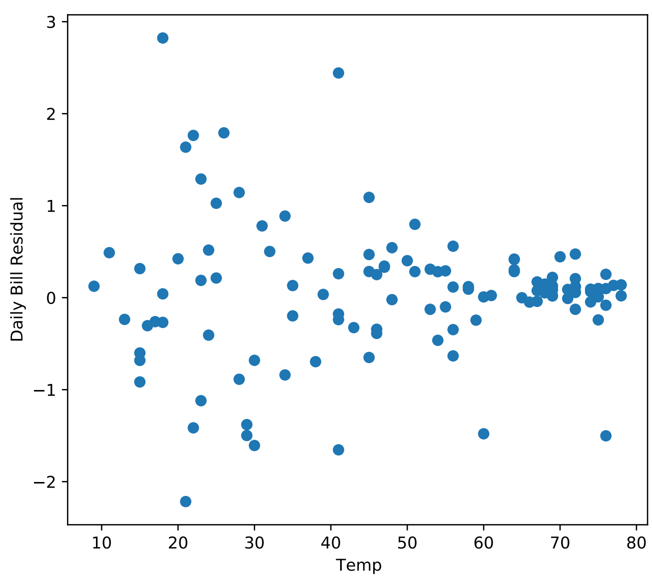



Figure 4: Residuals from Fitted Model with $h^{\star} = 6.9$.

controls the smoothness. Moreover, we can see that the covariance function $CSE$ is smoother than

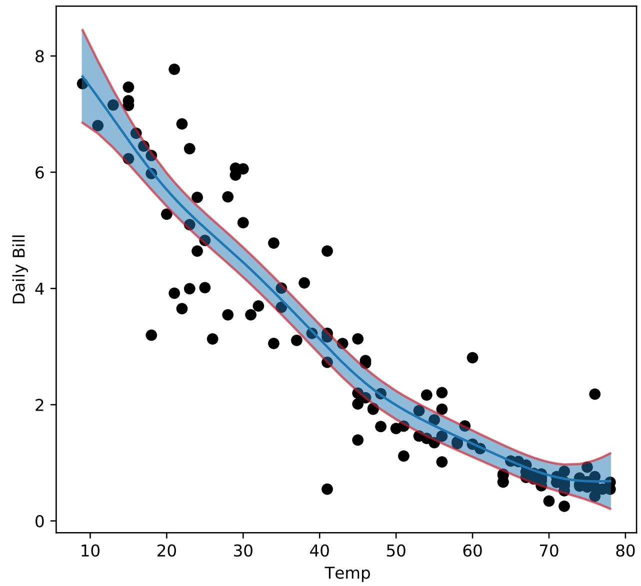



Figure 5: Confidence Interval from Fitted Model with $h^{\star} = 6.9$.

$CM52$.

(B) As derived in previous homework, for a multivariable Normal distribution:

$$\left[\begin{array}{c} y_1 \\ y_2 \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} \mu_1 \\ \mu_2 \end{array}\right], \left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array}\right]\right)$$

The conditional distribution is

$$p\left(y_1 \mid y_2\right) \sim \mathcal{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(y_2 - \mu_2\right), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

Let $\theta = \left(f\left(x_1\right), \ldots, f\left(x_n\right)\right)$ and $C$ is the given covariance function, the joint distribution of $\theta$ and $f(x^*)$ is

$$\left[\begin{array}{c} f\left(x^*\right) \\ \theta \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} m\left(x^*\right) \\ m(\mathbf{x}) \end{array}\right], \left[\begin{array}{cc} C^* & \tilde{C}^T \\ \tilde{C} & C \end{array}\right]\right)$$

where

$$C = C(\mathbf{x}, \mathbf{x}); \quad \tilde{C} = C\left(\mathbf{x}, x^*\right); \quad C^* = C\left(x^*, x^*\right)$$

Applying the conditional formular:

$$f\left(x^*\right) \mid \theta, \mathbf{x}, x^* \sim \mathcal{N}\left(m\left(x^*\right) + \tilde{C}^T C^{-1}(\theta - m(\mathbf{x})), C^* - \tilde{C}^T C^{-1}\tilde{C}\right)$$

(C) We have

$$
\begin{aligned}
p(\theta, y) &= p(y \mid \theta) \cdot p(\theta) \\
&\propto \exp\left\{-\frac{1}{2}[(y-R\theta)^T\Sigma^{-1}(y-R\theta)+(\theta-m)^T V^{-1}(\theta-m)]\right\} \\
&\propto \exp(A)
\end{aligned}
$$

We can rewrite $A$ as

$$
A \propto \left[\begin{array}{c} \theta-m \\ y-Rm \end{array}\right]^T \left[\begin{array}{cc} V^{-1}+R^T\Sigma^{-1}R & -R^T\Sigma^{-1} \\ -\Sigma^{-1}R & \Sigma^{-1} \end{array}\right] \left[\begin{array}{c} \theta-m \\ y-Rm \end{array}\right]
$$

In particular,

$$
\begin{aligned}
A \propto & (\theta-m)^T\left(V^{-1}+R^T\Sigma^{-1}R\right)(\theta-m)-(y-Rm)^T\Sigma^{-1}R(\theta-m) \\
& -(\theta-m)^T R^T\Sigma^{-1}(y-Rm)+(y-Rm)^T\Sigma^{-1}(y-Rm) \\
\propto & \theta^T\left(V^{-1}+R^T\Sigma^{-1}R\right)\theta-2m^T\left(V^{-1}+R^T\Sigma^{-1}R\right)\theta \\
& -y^T\Sigma^{-1}R\theta+y^T\Sigma^{-1}Rm+m^TR^T\Sigma^{-1}R\theta \\
& -\theta^TR^T\Sigma^{-1}y+\theta^TR^T\Sigma^{-1}Rm+m^TR^T\Sigma^{-1}y \\
& +y^T\Sigma^{-1}y^T-2m^TR^T\Sigma^{-1}y \\
= & \theta^TV^{-1}\theta+\theta^TR^T\Sigma^{-1}R\theta-2m^TV^{-1}\theta-2m^TR^T\Sigma^{-1}R\theta \\
& -y^T\Sigma^{-1}R\theta+y^T\Sigma^{-1}Rm+m^TR^T\Sigma^{-1}R\theta \\
& -\theta^TR^T\Sigma^{-1}y+\theta^TR^T\Sigma^{-1}Rm+m^TR^T\Sigma^{-1}y \\
& +y^T\Sigma^{-1}y^T-2m^TR^T\Sigma^{-1}y
\end{aligned}
$$

Therefore,

$$
p(\theta, y)=\mathcal{N}\left(\left[\begin{array}{c} \mathbf{m} \\ R\mathbf{m} \end{array}\right],\left[\begin{array}{cc} V^{-1}+R^T\Sigma^{-1}R & -R^T\Sigma^{-1} \\ -\Sigma^{-1}R & \Sigma^{-1} \end{array}\right]\right)
$$

**In nonparametric regression and Spatial Smoothing** :
(A) We denote $\theta=(f(x_1),\ldots,f(x_n))$

$$
y \sim \mathcal{N}\left(\theta, \sigma^2 I\right)
$$

Then

$$
\begin{aligned}
p(\theta \mid -) &\propto p\left(y \mid \theta, \sigma^2\right) p(\theta) \\
&\propto \exp\left(-\frac{1}{2}\left[(y-\theta)^T\left(\sigma^2 I\right)^{-1}(y-\theta)+\theta^T C^{-1}\theta\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left(-2y^T\left(\sigma^2 I\right)^{-1}\theta+\theta^T(\left(\sigma^2 I\right)^{-1}+C^{-1})\theta\right)\right) \\
&=\mathcal{N}\left(\left(I+\sigma^2C^{-1}\right)^{-1}y,\left(\sigma^{-2}I+C^{-1}\right)^{-1}\right)
\end{aligned}
$$

(B) Using the derived properties of joint and conditional distribution in the previous part, we have

$$
\begin{aligned}
E\left[f\left(x^{*}\right) \mid y\right] &= m\left(x^{*}\right)+\tilde{C}^{T}(C+\sigma^{2} I)^{-1}(y-m(\mathbf{x})) \\
&= \tilde{C}^{T}(C+\sigma^{2} I)^{-1} y \qquad (m(x^{*})=0)) \\
&= \sum_{i=1}^{n} \tilde{C}\left(x_{i}, x^{*}\right)\left(C+\sigma^{2} I\right)^{-1}\left(x_{i}, x_{i}\right) y_{i} \\
&= \sum_{i=1}^{n} w_{i} y_{i}
\end{aligned}
$$

Similarly, we can write

$$
\begin{aligned}
E\left[f\left(x^{*}\right) \mid y\right] = \sum_{i=1}^{n} \alpha_{i} C\left(x_{i}, x^{*}\right), \\
\alpha_{i} = (C+\sigma^{2} I)^{-1} y_{i}
\end{aligned}
$$

For the variance, we have

$$
Var[f(x*)] = C^{*} - \tilde{C}^{T}(C+\sigma^{2} I)^{-1} \tilde{C}
$$

where

$$
C = C(\mathbf{x}, \mathbf{x}); \quad \tilde{C} = C\left(\mathbf{x}, x^{*}\right); \quad C^{*} = C\left(x^{*}, x^{*}\right)
$$

(C) I set $\sigma^2 = 0.61$, $b \in \{1, 5, 10\}$, $\tau_1 \in \{1, 5, 10\}$, and $\tau_2 = 1e-6$. I plot estimated functions and the corresponding 95% confidence interval in Figure **??**. We observe that, $b$ increases leading to smoother functions.
(D) Using the previous part, we have

$$
p(y) = \mathcal{N}\left(0, C+\sigma^{2} I\right)
$$

(E) We have the likelihood function

$$
\begin{aligned}
\log p(y) &= \log \left(\left|2 \pi\left(C+\sigma^{2} I\right)\right|^{-1 / 2} \exp \left(-\frac{1}{2} y^{T}\left(C+\sigma^{2} I\right)^{-1} y\right)\right) \\
&= -\frac{1}{2} \log \left|2 \pi\left(C+\sigma^{2} I\right)\right| - \frac{1}{2} y^{T}\left(C+\sigma^{2} I\right)^{-1} y \\
&= -\frac{n}{2} \log (2 \pi) - \frac{1}{2} \log \left|C+\sigma^{2} I\right| - \frac{1}{2} y^{T}\left(C+\sigma^{2} I\right)^{-1} y
\end{aligned}
$$

I search for best $\tau_1$ and $b$ on a grid of $100 \times 100$ values between $[0.1, 100]$. I got $b^\star = 57.93$ and $\tau_1^\star = 5.35$. I show the estimated functions and confidence interval in Figure 10.

(F) For keeping simplicity, I choose the Eculidean distance. I search for best hyper-parameters like in the previous part. I plot scatter-plots of data, predicted posterior means, predicted posterior variances in Figure 11.

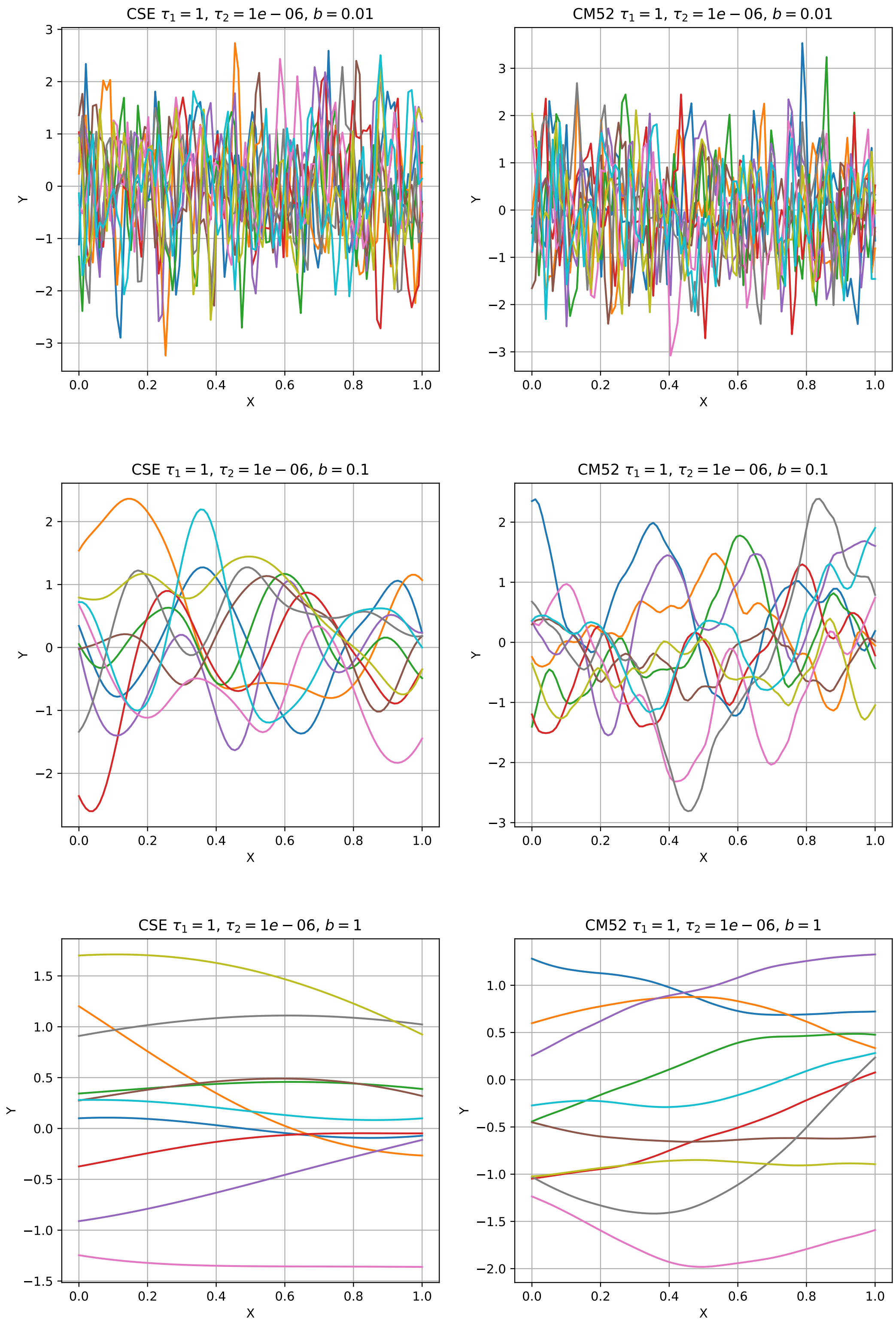



Figure 6: 10 random samples from GP with different $b$.

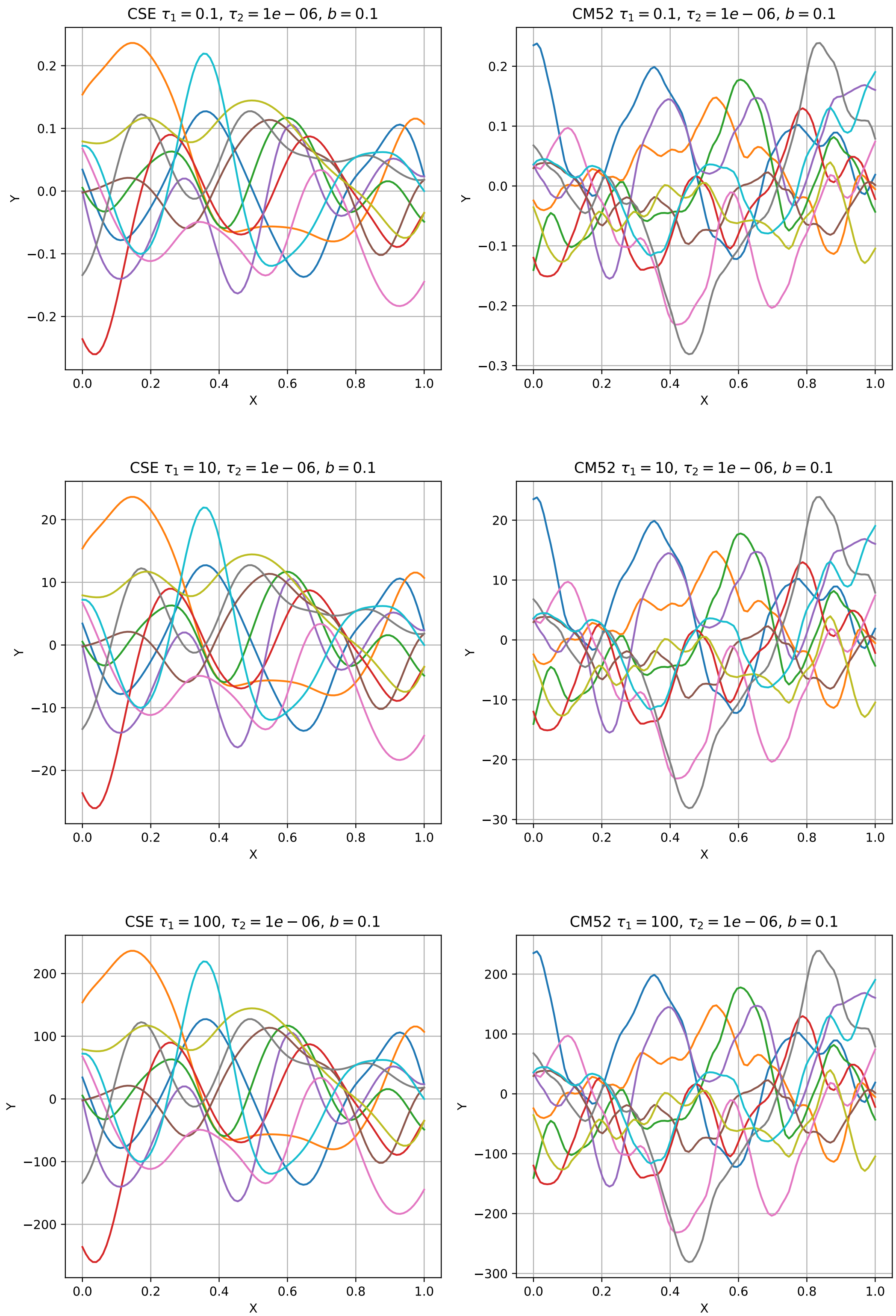



Figure 7: 10 random samples from GP with different $\tau_1$.


Figure 8: 10 random samples from GP with different $\tau_2$.

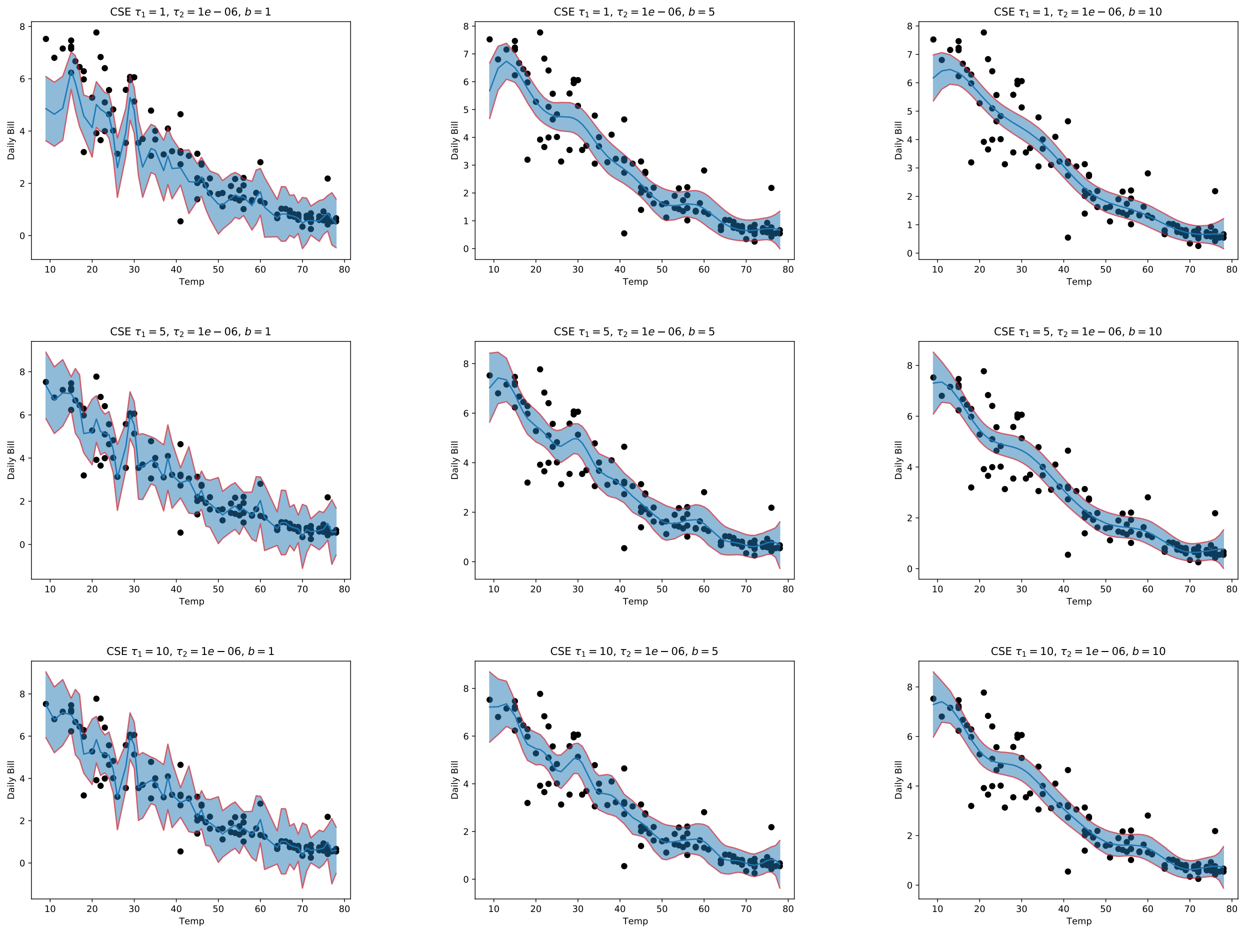



Figure 9: Gaussian Processes for Utilities Data.

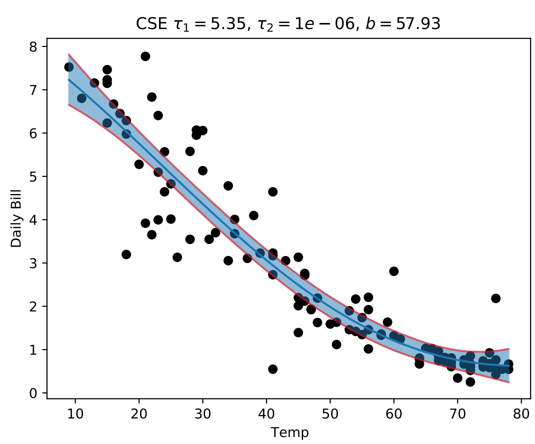



Figure 10: Gaussian Processes for Utilities Data with best choice of $\tau_1$ and $b$.

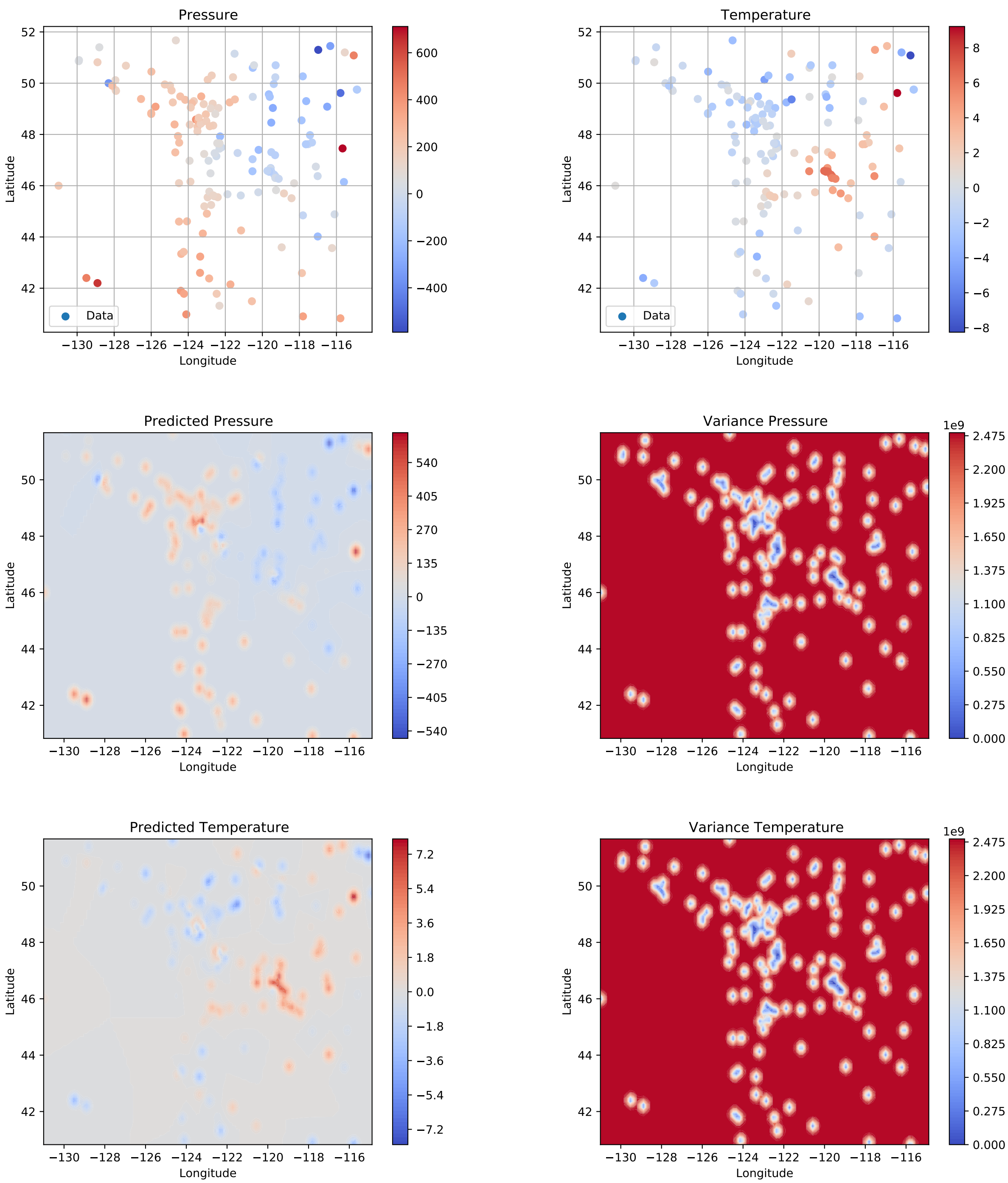



Figure 11: Gaussian Processes for Weather Data.