Homework 6

Solutions

1. (Gradient descent and ridge regression)

   (a) The gradient equals

$$\nabla f(\beta) = (XX^T + \lambda I)\beta - Xy. \tag{1}$$

   The gradient-descent updates are

$$\beta^{(k+1)} = \left(I - \alpha(XX^T + \lambda I)\right)\beta^{(k)} + \alpha Xy, \tag{2}$$

   which yields

$$\beta^{(k)} = \sum_{i=0}^{k-1} \left(I - \alpha(XX^T + \lambda I)\right)^i \alpha Xy. \tag{3}$$

   Since $p \leq n$ and X is full rank, we have $UU^T = U^TU = I$, so that

$$\beta^{(k)} = \alpha \sum_{i=0}^{k-1} \left(UU^T - \alpha(US^2U^T + \lambda UU^T)\right)^i USV^Ty \tag{4}$$

$$= \alpha U \sum_{i=0}^{k} \left(I - \alpha(S^2 + \lambda I)\right)^i SV^Ty \tag{5}$$

$$= \alpha U \operatorname{diag}_{j=1}^{p} \left(\sum_{i=0}^{k} \left(1 - \alpha(s_j^2 + \lambda)\right)^i \right) SV^Ty. \tag{6}$$

   By the geometric-sum formula we conclude:

$$\beta^{(k)} = \alpha U \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha(s_j^2 + \lambda)\right)^k}{\alpha(s_j^2 + \lambda)} \right) SV^Ty. \tag{7}$$

   (b) We need $\lim_{k\to\infty} \left(1 - \alpha(s_j^2 + \lambda)\right)^k = 0$, which happens if $0 < \alpha < 2/(s_1^2 + \lambda) \leq 2/(s_j^2 + \lambda)$ for $1 \leq j \leq p$.

   (c) To ease notation, let $\tau_j := 1 - \alpha(s_j^2 + \lambda)$. By the answer to the first question

$$\tilde{\beta}^{(k)} = U \operatorname{diag}_{j=1}^{p} \left(\frac{s_j(1 - \tau_j^k)}{s_j^2 + \lambda} \right) V^T \left(X^T\beta_{\text{true}} + \tilde{z}_{\text{train}} \right) \tag{8}$$

$$= U \operatorname{diag}_{j=1}^{p} \left(\frac{s_j(1 - \tau_j^k)}{s_j^2 + \lambda} \right) V^T \left(VSU^T\beta_{\text{true}} + \tilde{z}_{\text{train}} \right) \tag{9}$$

$$= U \operatorname{diag}_{j=1}^{p} \left(\frac{s_j^2(1 - \tau_j^k)}{s_j^2 + \lambda} \right) U^T\beta_{\text{true}} + U \operatorname{diag}_{j=1}^{p} \left(\frac{s_j(1 - \tau_j^k)}{s_j^2 + \lambda} \right) V^T\tilde{z}_{\text{train}}. \tag{10}$$

By Theorem 8.6 in the PCA lecture notes $\tilde{\beta}^{(k)}$ is a Gaussian random vector with mean

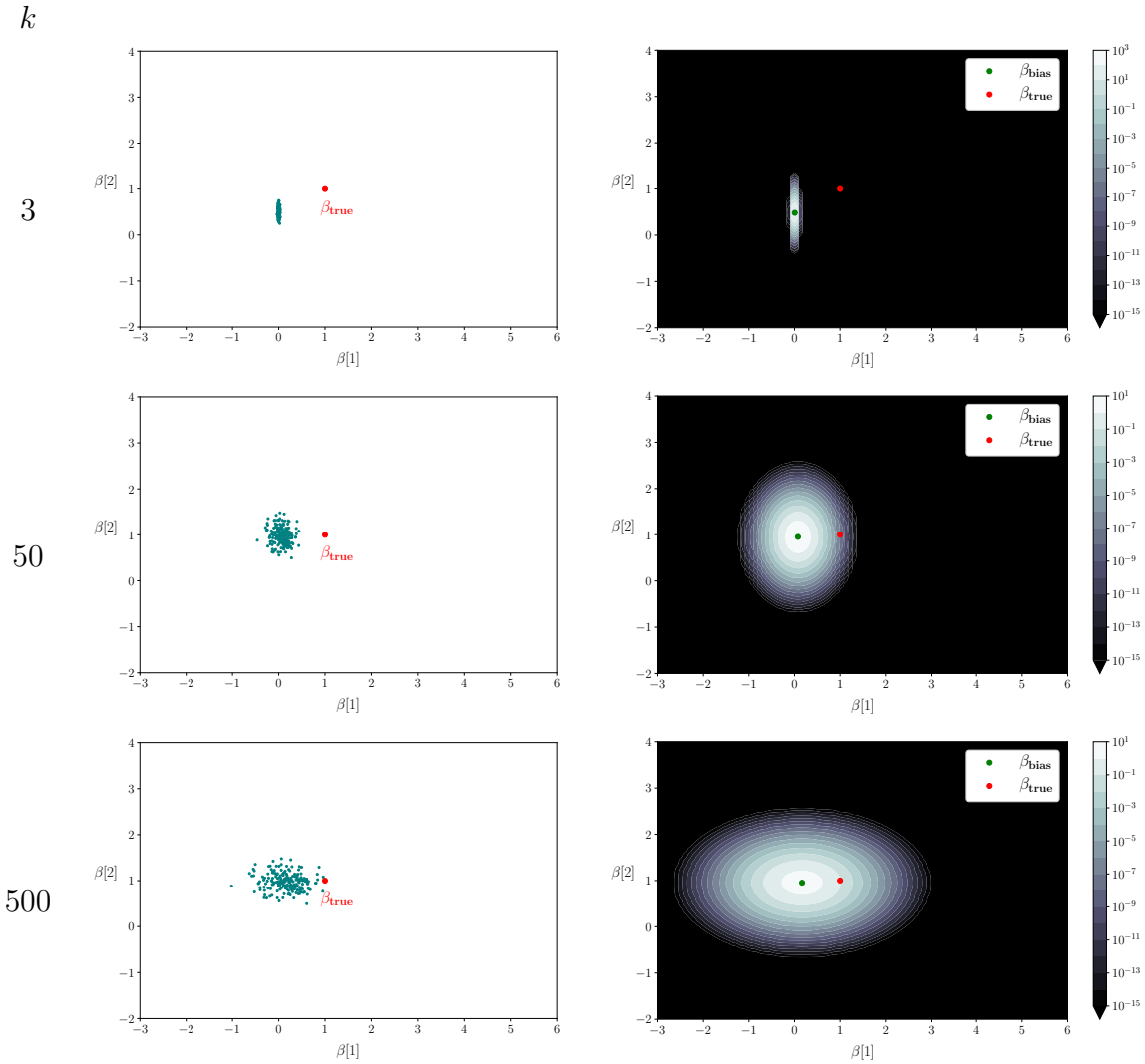
$$\beta_{\text{bias}} := \sum_{j=1}^{p} \left(\frac{s_j^2(1-(1-\alpha(s_j^2+\lambda))^k)}{s_j^2+\lambda} \right) \langle u_j, \beta_{\text{true}} \rangle \, u_j \tag{11}$$

and covariance matrix

$$\Sigma_{\text{GD}} := \sigma^2 U \operatorname{diag}_{j=1}^{p} \left(\left(\frac{s_j(1-(1-\alpha(s_j^2+\lambda))^k)}{s_j^2+\lambda} \right)^2 \right) U^T, \tag{12}$$

where $\operatorname{diag}_{j=1}^{p}(d_i)$ denotes a diagonal matrix with entries $d_1, \ldots, d_p$.

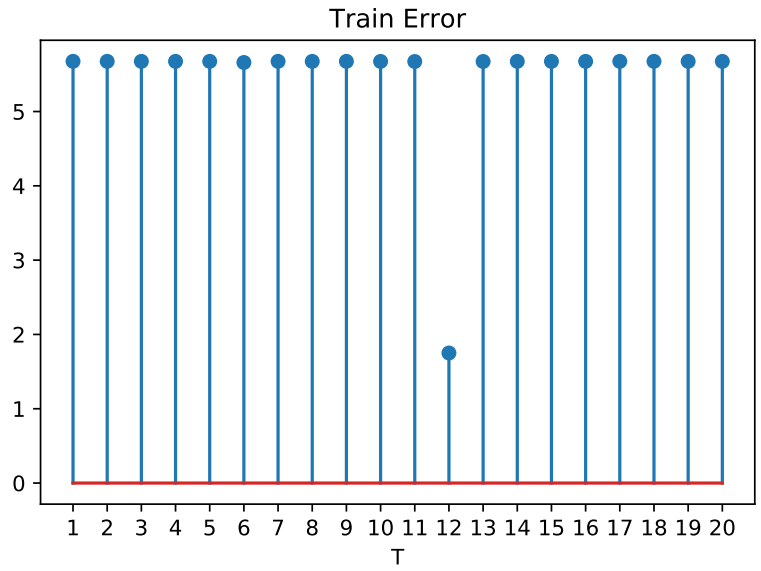
(d) The figures are the following:

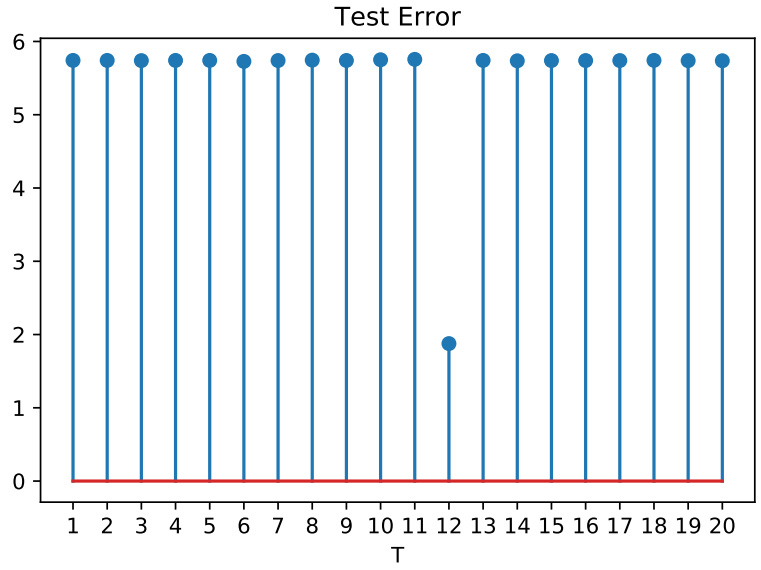


2. (Climate modeling)

(a) We have four parameters in the model and 1800 data points to fit the model. The model is very unlikely to overfit.

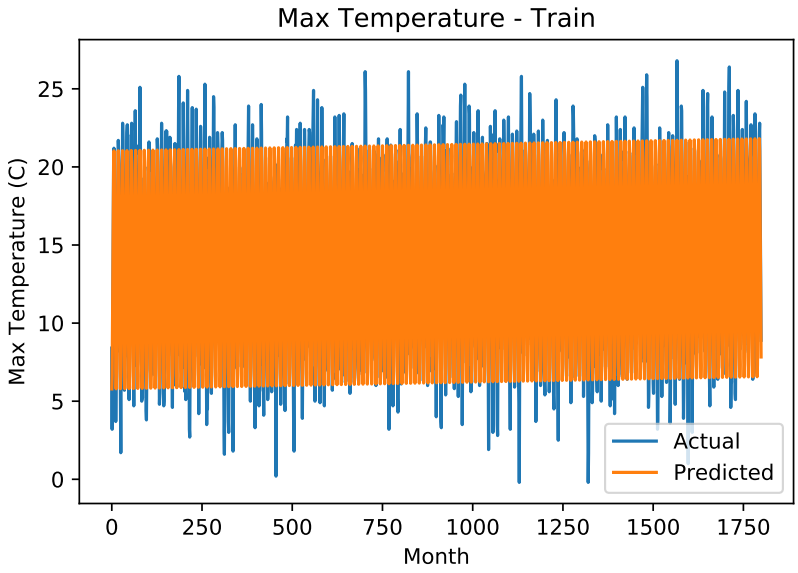
(b) Below we show a plot of the test and train error with the period T:.

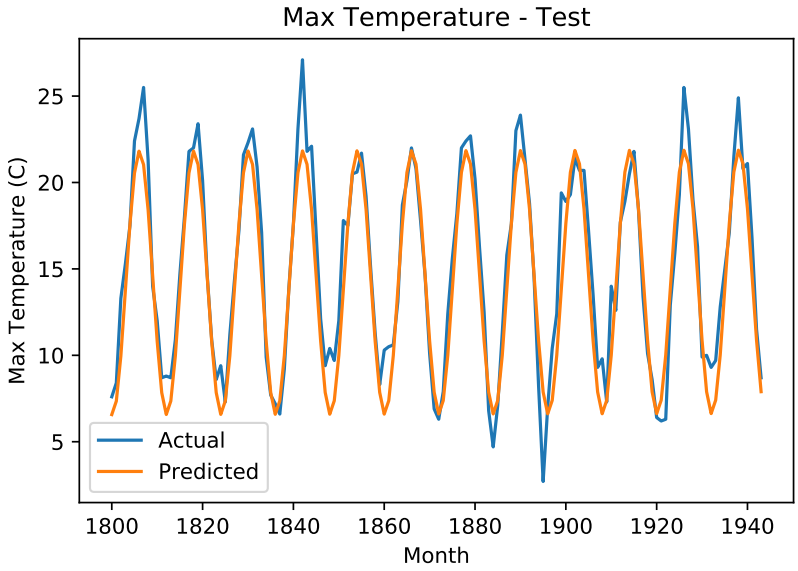




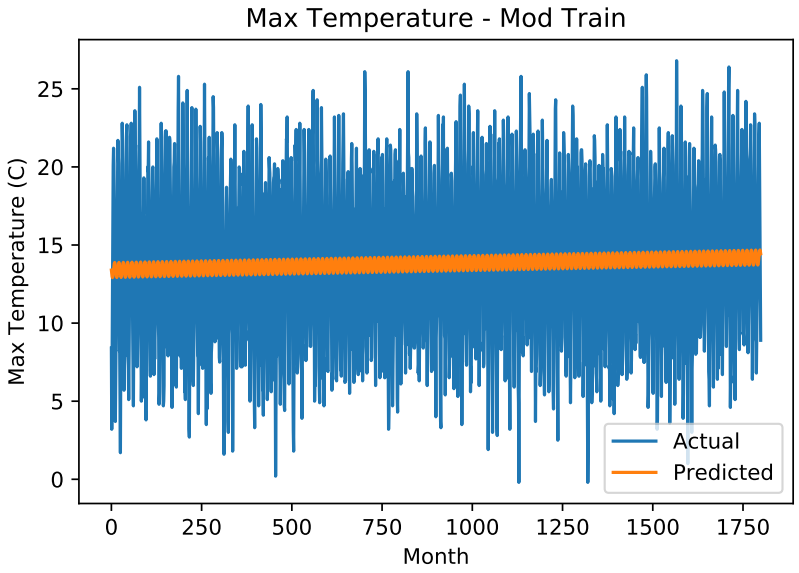
$T^* = 12$. This is not surprising, since we expect weather to be cyclical with a 12 month period.

(c) We show the curve we fit overlayed on top of the data below:





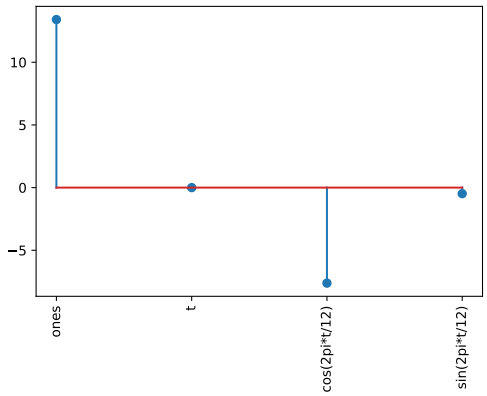
(d) If we remove the term with cosine, we obtain the following fit for train data:



On the training set, this fit has an RMSE error of 5.665 as compared to the RMSE error of 1.75 when cosine term is also included. This is fit is bad because the phases don't align when you only include the sine term. You need the cosine term as well so that you can fit any phase.

(e) For completeness, the full set of coefficients we obtained are

component	values
1	$1.34 \cdot 10^{1}$
t	$4.38 \cdot 10^{-4}$
$\cos(2\pi t/12)$	$-7.61 \cdot 10^{0}$
$\sin(2\pi t/12)$	$-4.83 \cdot 10^{-1}$



The ones account for the bias term or the offset. t term captures any linear trend we have in the data. The cos and sin term together accounts for periodic trend in data. On combining $c_1 cos(t) + c_2 sin(t)$ one can create $c\ cos(t + \theta)$ which models a sinusoid with arbitrary phase.

Since the coefficent of t term in our model is positive, the temperature in oxford is increasing on an average. In degrees per century, the temperature trend is $+0.5225$.

3. (Sines and cosines)

(a) Note that

$$\hat{x}[k] = \int_{-1/2}^{1/2} x(t)e^{-2\pi ikt}\,dt = \int_{-1/2}^{1/2} x(t)\overline{e^{-2\pi i(-k)t}}\,dt = \overline{\int_{-1/2}^{1/2} x(t)e^{-2\pi i(-k)t}\,dt} = \overline{\hat{x}[-k]}.$$

(b) By definition we have

$$\mathcal{F}_{k_c}x(t) = \sum_{k=-k_c}^{k_c} \hat{x}[k]e^{2\pi ikt}.$$

We first let $a_0 = \hat{x}[0]$. Note that $\overline{a_0} = \overline{\hat{x}[0]}$ by the previous part, so $a_0 \in \mathbb{R}$. Grouping positive and negative terms we have

$$\begin{aligned}
\hat{x}[k]e^{2\pi ikt} + \hat{x}[-k]e^{2\pi i(-k)t} &= \hat{x}[k]e^{2\pi ikt} + \overline{\hat{x}[k]e^{2\pi ikt}} \quad \text{(by previous part)} && (13)\\
&= 2\Re\left(\hat{x}[k]e^{2\pi ikt}\right) && (14)\\
&= 2\Re(\hat{x}[k])\cos(2\pi kt) - 2\Im(\hat{x}k)\sin(2\pi kt). && (15)
\end{aligned}$$

This proves the given series formulation where $a_k = 2\Re(\hat{x}[k])$ and $b_k = -2\Im(\hat{x}[k])$ for $k \geq 1$.

(c) By definition, the expressions from the previous part are given by

$$a_k = 2\int_{-1/2}^{1/2} x(t)\cos(2\pi kt)\,dt \quad \text{and} \quad b_k = 2\int_{-1/2}^{1/2} x(t)\sin(2\pi kt)\,dt.$$

They are inner products between x and cosine and sine functions respectively.

(d) Note that

$$\cos(2\pi(t+\phi)) = \frac{e^{2\pi i(t+\phi)} + e^{-2\pi i(t+\phi)}}{2} = \frac{e^{2\pi i\phi}}{2}e^{2\pi it} + \frac{e^{-2\pi i\phi}}{2}e^{-2\pi it}.$$

Thus we have

$$\hat{x}[1] = \frac{e^{2\pi i\phi}}{2}, \quad \hat{x}[-1] = \frac{e^{-2\pi i\phi}}{2},$$

and $\hat{x}[k] = 0$ for $|k| \neq 1$.

(e) We will prove that $\overline{\hat{x}[k]} = \hat{x}[k]$ for all $k \in \mathbb{Z}$. By the first part we have $\hat{x}[0] = \overline{\hat{x}[0]}$. For $k \neq 0$ we have

$$\begin{aligned}
\hat{x}[k] &= \int_{-1/2}^{1/2} x(t)e^{-2\pi ikt}\,dt && (16)\\
&= \int_{1/2}^{-1/2} -x(-u)e^{2\pi iku}\,du \qquad (u = -t) && (17)\\
&= \int_{-1/2}^{1/2} x(-u)e^{2\pi iku}\,du && (18)\\
&= \int_{-1/2}^{1/2} x(u)e^{2\pi iku}\,du \qquad (x \text{ is even}) && (19)\\
&= \hat{x}[-k]. && (20)
\end{aligned}$$

So by the first part this gives

$$\hat{x}[k] = \hat{x}[-k] = \overline{\hat{x}[k]}$$

as required.