תרגיל 3 | BML
מגיש: אבי כוגן | ת.ז: 205417710

שאלה 1

$y(x) = \Theta^T h(x) + \eta$

כאשר $h : R^d \rightarrow R^p, \Theta \in R^{p\times k}, \eta \sim N(0, \sigma^2 I)$ לכן $y \in R^k$.
נגדיר $Y \in R^{N\times k}$ מטריצת הפלטים של על N נק' אימון.
מתקיים

$$logp(y|\Theta, \sigma) = \Sigma_{i=1}^N log\mathcal{N}(y_i|\Theta^T h(x_i), I\sigma^2)$$
$$=$$

נמצא את הMLE עבור Θ:

$$\Sigma_{i=1}^N log\mathcal{N}(y_i|\Theta^T h(x_i), I\sigma^2) =$$
$$\Sigma_{i=1}^N log(2\pi)^{-\frac{k}{2}} \cdot det(\sigma^2 I)^{-\frac{1}{2}} \cdot e^{-\frac{||\Theta^T h(x_i) - y_i||^2}{2\sigma^2}} =$$
$$\Sigma_{i=1}^N log(2\pi)^{-\frac{k}{2}} \sigma^{-\frac{k}{2}} e^{-\frac{||\Theta^T h(x_i) - y_i||^2}{2\sigma^2}} =$$
$$-\frac{Nk}{2} log(2\pi) - \frac{Nk}{2} log\sigma + \Sigma_{i=1}^N - \frac{||y_i - \Theta^T h(x_i)||^2}{2\sigma^2}$$

$$\frac{\partial}{\partial\Theta} \Sigma_{i=1}^N log\mathcal{N}(y_i|\Theta^T h(x_i), I\sigma^2)$$
$$= \Sigma_{i=1}^N - \frac{1}{\cancel{2}\sigma^2} \cancel{2}(y_i - \Theta^T h(x_i))h(x_i)^T$$

נשווה ל-0:

$$\Sigma_{i=1}^{N}\cancel{\frac{1}{\sigma^2}}(y_i-\Theta^T h(x_i))\cdot\cancel{}h(x_i)^T \overset{!}{=} 0$$
$$\Leftrightarrow \Sigma_{i=1}^{N} y_i h(x_i)^T - \Theta^T\Sigma_{i=1}^{N} h(x_i)h(x_i)^T = 0$$
$$\Leftrightarrow \Sigma_{i=1}^{N} h(x_i)y_i - (\Sigma_{i=1}^{N} h(x_i)h(x_i)^T)\cdot\Theta = 0$$
$$\Leftrightarrow \Sigma_{i=1}^{N} h(x_i)y_i = (\Sigma_{i=1}^{N} h(x_i)h(x_i)^T)\cdot\Theta$$
$$*assumes\ that\ the\ symmetric\ matrix \Sigma_{i=1}^{N} h(x_i)h(x_i)^T \in R^{p\times p}\ invertble$$
$$(\Sigma_{j=1}^{N} h(x_j)h(x_j)^T)^{-1}\Sigma_{i=1}^{N} y_i h(x_i)^T = \Theta^T$$

$$\Rightarrow \hat{\Theta}^{MLE} = (\Sigma_{j=1}^{N} h(x_j)h(x_j)^T)^{-1}\Sigma_{i=1}^{N} h(x_i)y_i^T$$

אם נסמן: $H=\begin{pmatrix} - & h(x_1)^T & - \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ - & h(x_N)^T & - \end{pmatrix}\in R^{N\times p}$ נקבל: $\hat{\Theta}^{MLE}=(H^TH)^{-1}H^TY$.

שאלה 2

נסמן את העמודה ה-k של $\hat{\Theta}^{MLE}$ בתור $\hat{\Theta}_k^{MLE}$, נסמן את העמודה הk של Y בתור Y_k, לכן מכפל מטריצות מתקיים:

$$\hat{\Theta}_k^{MLE} = [(H^TH)^{-1}H^T]\cdot Y_k$$

קיבלנו שכל עמודה ב-$\hat{\Theta}_k^{MLE}$ ניתן לקביעה ע"י הפלט Y_k באופן ב"ת משאר הפלטים.
לכן ניתן לקבוע אומד ML לכל עמודה ב-Θ לפי כל פלט בנפרד ולשרשר אותם אחד אחרי השני לקבלת המודל המוכלל עם מספר פלטים.

שאלה 3

מתקיים K היא PSD לכן קיימת מטריצה כלשהי R כך ש $K=R^TR$.
לכן נקבל $k(x_i,x_j)=K_{i,j}=(R^T)_iR^j=(R^i)^TR^j$ מהגדרת כפל מטריצות, כאשר נסמן R_i השורה ה-i של R וR^j העמודה ה-j של R.
לכן אם נגדיר $f(x_i)=R^i$ נקבל שמתקיים $k(x_i,x_j)=f(x_i)^Tf(x_j)$.
לכן אם נגדיר בהתאם את f ו g עבור k_1 ו-k_2 בהתאמה ונסמן f_i,g_i האינדקס ה-i בפלט שלהן, נ

$$k_1(x,y)k_2(x,y) = f^T(x)f(y)g^T(x)g(y)$$
$$= \Sigma_i\Sigma_j g_i^T(x)f_j^T(x)f_j(y)g_i(y)$$

שאלה 4

ממשפט מרסר מתקיים:

$$\begin{aligned} k_1(x,y) =& a(x)^T a(y) \ s.t \ a(x) = (a_1(x),..,a_m(x)) \\ k_2(x,y) =& b(x)^T b(y) \ s.t \ b(x) = (a_1(x),..,a_n(x)) \end{aligned}$$

לכן נקבל:

$$\begin{aligned} k_\times(x,y) &= k_1(x,y)k_2(x,y) \\ &= [\Sigma_{i=1}^n a_i(x)a_i(y)] \cdot [\Sigma_{j=1}^n b_j(x)b_j(y)] \\ &= \Sigma_{i=1}^n \Sigma_{n=1}^m a_i(x)a_i(y)b_j(x)b_j(y) \\ &= \Sigma_{i=1}^n \Sigma_{n=1}^m a_i(x)b_j(x)a_i(y)b_j(y) \\ &= c^T(x)c(x) \end{aligned}$$

כאשר $c_{i\cdot j}(x) = a_i(x)b_j(x)$ $R^{mn} \ni$ $i \in [m], j \in [n]$.
קיבלנו ש- $k_\times$ הוא מכפלה פנימית של שני וקטורים אזי הוא $valid\ kernel$ לפי טענה שראינו בתרגול.

שאלה 5

לא $valid\ kernel$, דוגמה נגדית:
מתקיים:

$$K = \begin{pmatrix} e^{||x-x||^2} & e^{||x-y||^2} \\ e^{||y-x||^2} & e^{||y-y||^2} \end{pmatrix} = \begin{pmatrix} 1 & e^{||x-y||^2} \\ e^{||y-x||^2} & 1 \end{pmatrix}$$

K הוא PSD אם"ם $det(K) \geq 0$ כלומר:

$$\begin{aligned} |K| \geq& 0 \Leftrightarrow \\ 1 - e^{2||x-y||^2} \geq& 0 \Leftrightarrow \\ 1 \geq& e^{2||x-y||^2} \Leftrightarrow \\ 0 \geq& 2||x-y||^2 \Leftrightarrow \\ 0 \geq& ||x-y||^2 \end{aligned}$$

כלומר קיבלנו ש-K PSD אם"ם $0 \geq ||x-y||^2$, לכן עבור x,y המקיימים $||x-y||^2 > 0$, כלומר לכל $x \neq y$ מתקיים k לא $valid\ kernel$.

שאלה 6

מתקיים:

$$k(x,y) = k_1(x,y) - k_2(x,y)$$

ה-$kernel$ לא חוקי, דוג' נגדית:
עבור $\mathcal{X} = R^2$,
נסמן ב-$K^{(i)}$ את מטריצת הגראם של $k_i(x,y)$ וב-K את מטריצת הגראם של $k(x,y)$, אזי מתקיים:

$$K = K^{(1)} - K^{(2)}$$

הטענה היא לכל $valid\ kernel$,
נגדיר $k_1(x,y) = 0$ לכן $K^{(1)} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ היא PSD.
נגדיר $k_1(x,y) = 1$ לכן $K^{(2)} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ גם כן PSD.
לכן 2 ה-$kernl$-ים הם תקינים.
אבל מתקיים שעבור $v = (1,0)^T$:

$$\begin{aligned} v^T K v &= v^T K^{(1)} v - v^T K^{(2)} v \\ &= 0 - 1 = -1 < 0 \end{aligned}$$

לכן K אינה PSD.

שאלה 7

מתקיים ה$kernel$ חוקי.
הוכחה:
נסמן בK_a, K_b את מטריצות הגראם של k_a, k_b בהתאמה.

$$K_{i,j} = k(x^{(i)}, x^{(j)}) = k_a(x_a^{(i)}, x_a^{(j)}) + k_b(x_b^{(i)}, x_b^{(j)})$$

לכן נקבל:

$$K = K_a + K_b$$

ונקבל שלכל v מתקיים:

$$v^T K v = v^T K_a v + v^T K_b v \overset{*}{\geq} 0$$

* מכיוון שנתון K_a, K_b הם PSD.

שאלה 8

נפתר בתרגול:
מתקיים:

$$y|k(\cdot,\cdot),\sigma^2 = K\alpha + \begin{pmatrix}\eta \\ . \\ \eta\end{pmatrix}$$

לכן $\alpha \sim N$, $\begin{pmatrix}\eta \\ . \\ \eta\end{pmatrix} \sim N$,$y|k(\cdot,\cdot),\sigma^2$ נמצא את הפרטרים של ההתפלגות:

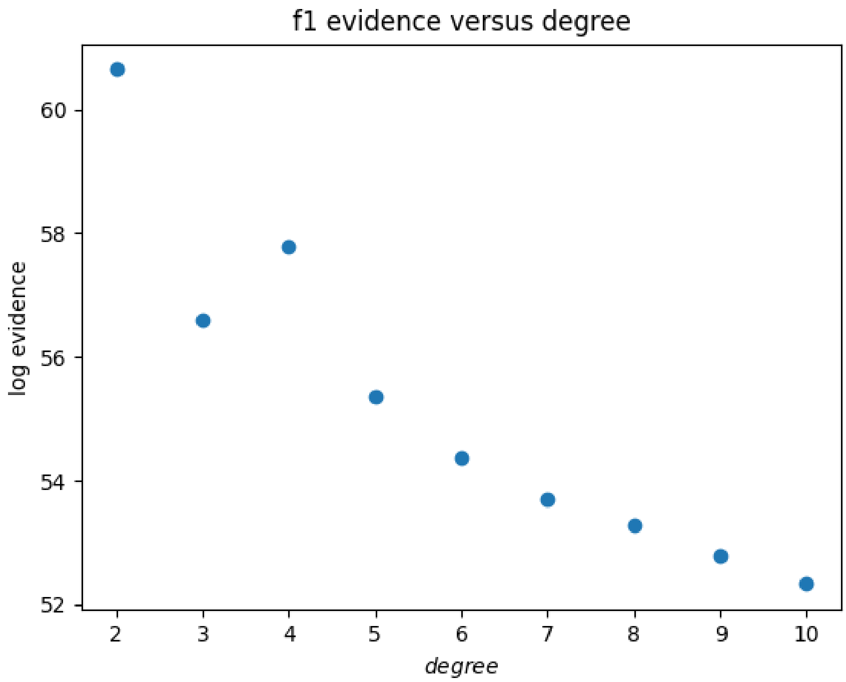
$$E[y|k(\cdot,\cdot),\sigma^2] = E[K\alpha + \begin{pmatrix}\eta \\ . \\ \eta\end{pmatrix}] = 0$$

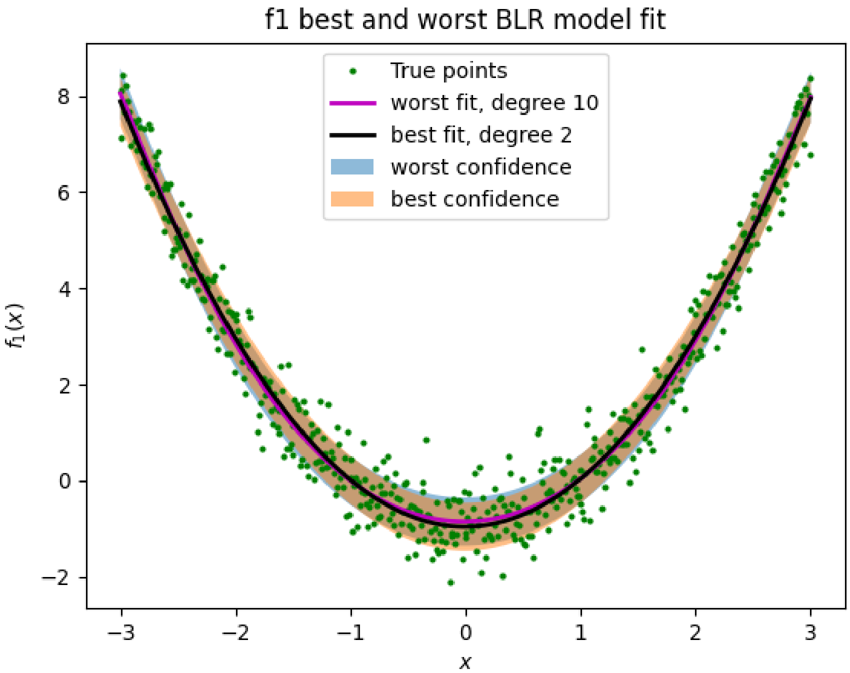
$$\begin{aligned} Cov[y|k(\cdot,\cdot),\sigma^2] &= Cov[K\alpha + \begin{pmatrix}\eta \\ . \\ \eta\end{pmatrix}] \\ &\overset{\alpha\perp\eta, K=K^T}{=} KCov[\alpha]K + Cov[\begin{pmatrix}\eta \\ . \\ \eta\end{pmatrix}] \\ &= KK^{-1}K + \sigma^2 I = K + \sigma^2 I \end{aligned}$$

לכן, $y|k(\cdot,\cdot),\sigma^2 \sim N(0, K+\sigma^2 I)$ כלומר $p(y|k(\cdot,\cdot),\sigma^2) = \mathcal{N}(y|0, K+\sigma^2 I)$.

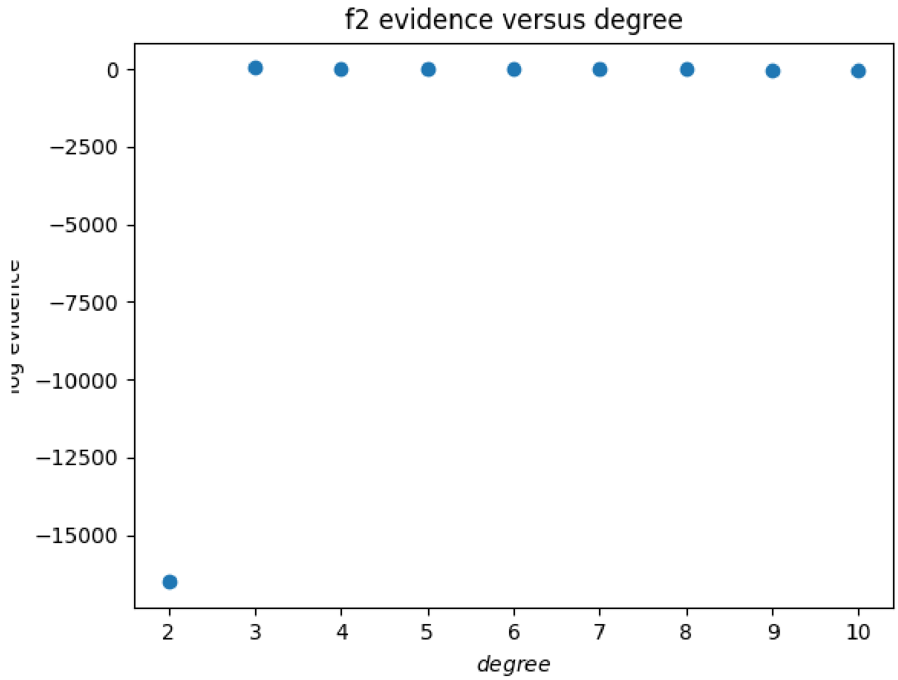
חלק פרקטי

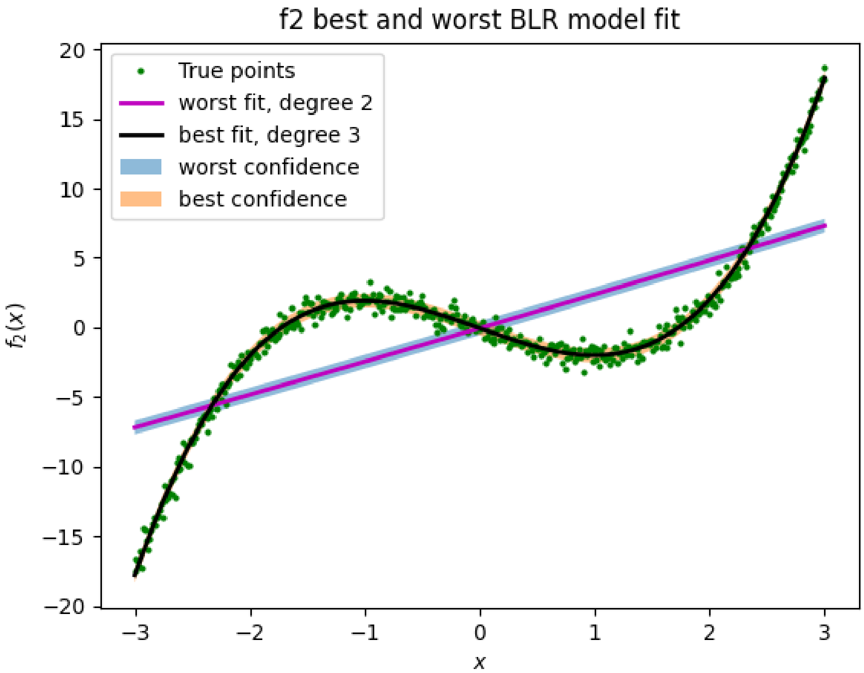
חלק 2.1

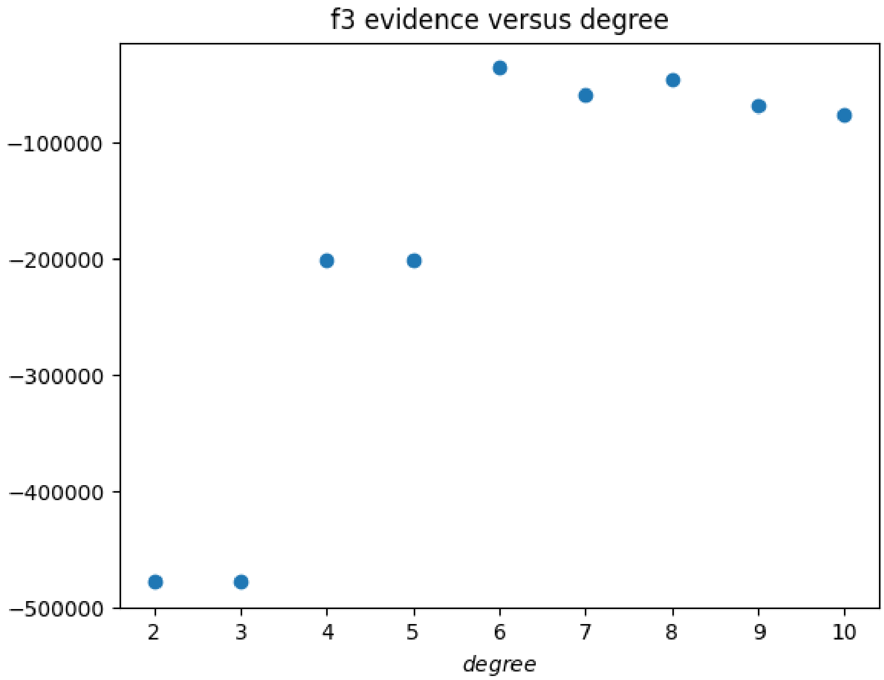


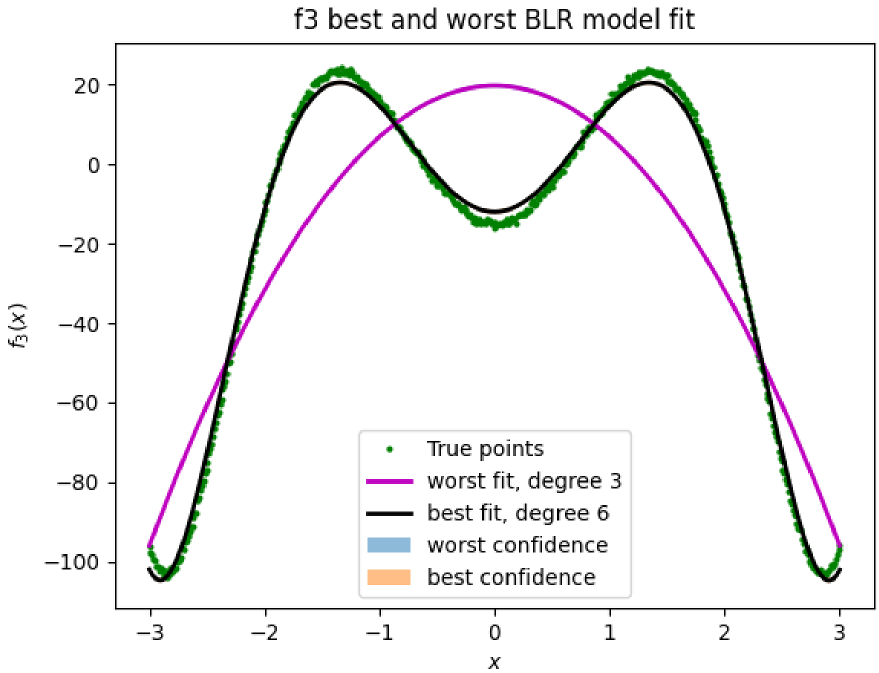


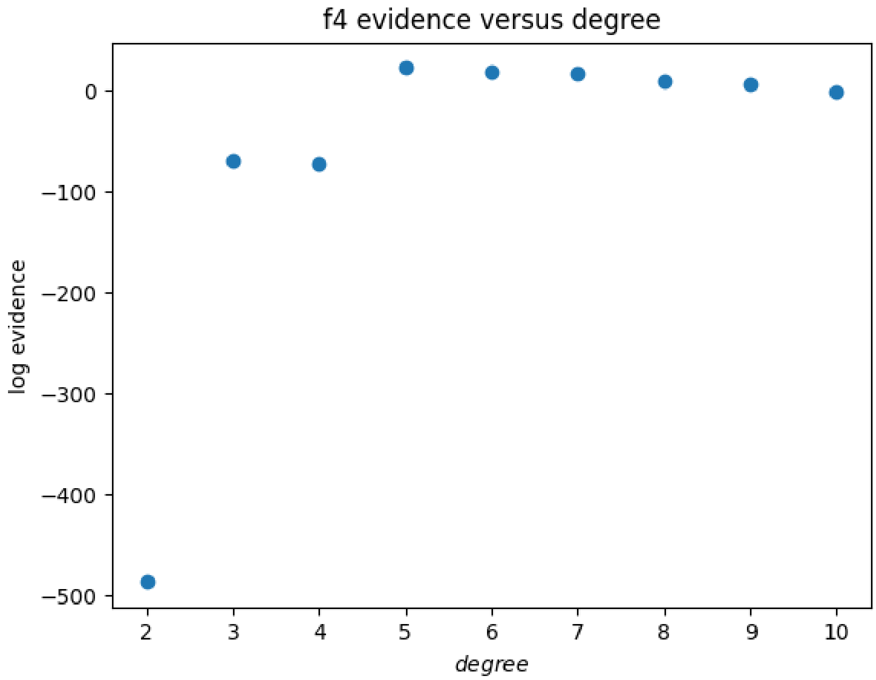
f1 best and worst BLR model fit
True points
worst fit, degree 10
best fit, degree 2
worst confidence
best confidence
8
6
4
2
0
−2
$f_1(x)$
−3
−2
−1
0
1
2
3
x

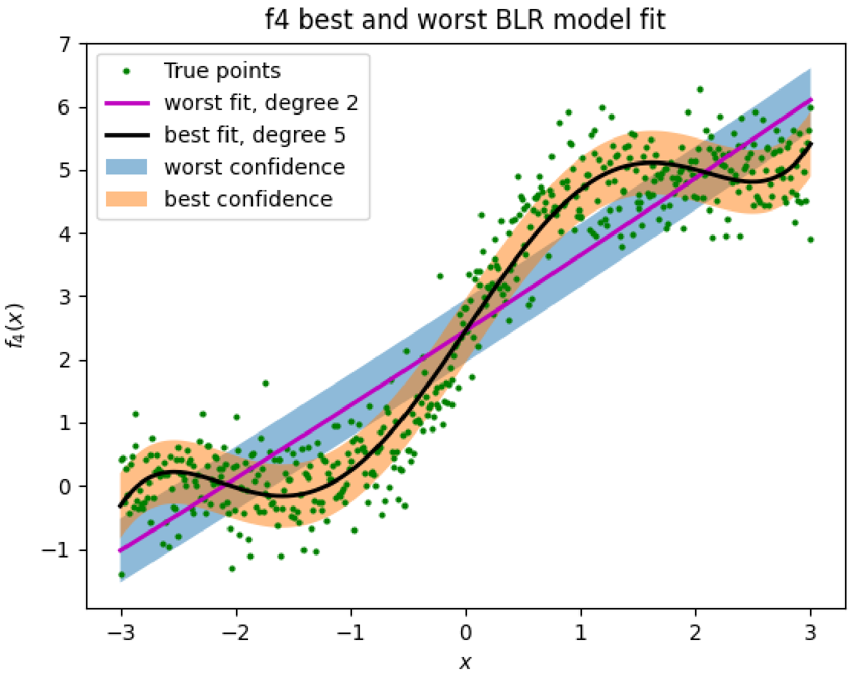

f2 evidence versus degree
0
−2500
−5000
−7500
−10000
−12500
−15000
2
3
4
5
6
7
8
9
10
degree

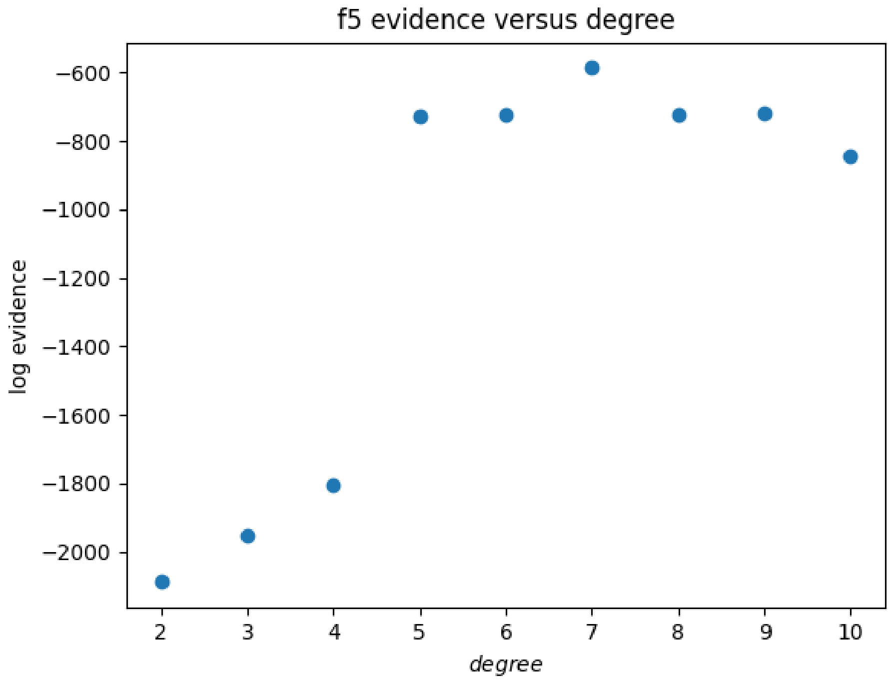

f2 best and worst BLR model fit
True points
worst fit, degree 2
best fit, degree 3
worst confidence
best confidence
$f_2(x)$
x
20
15
10
5
0
−5
−10
−15
−20
−3
−2
−1
0
1
2
3

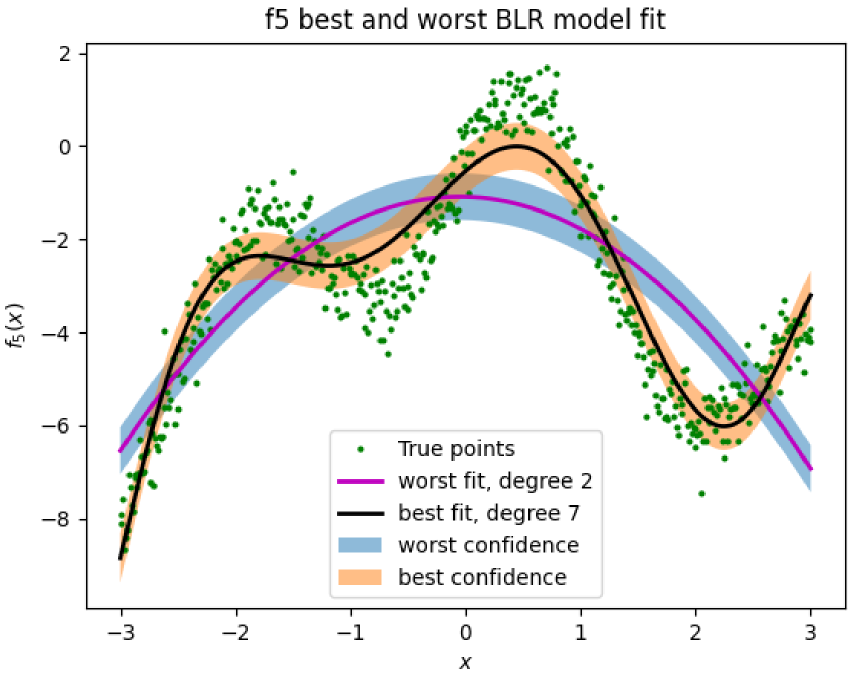

f3 evidence versus degree
−100000
−200000
−300000
−400000
−500000
2
3
4
5
6
7
8
9
10
degree


f3 best and worst BLR model fit
True points
worst fit, degree 3
best fit, degree 6
worst confidence
best confidence
$f_3(x)$
x
20
0
−20
−40
−60
−80
−100
−3
−2
−1
0
1
2
3

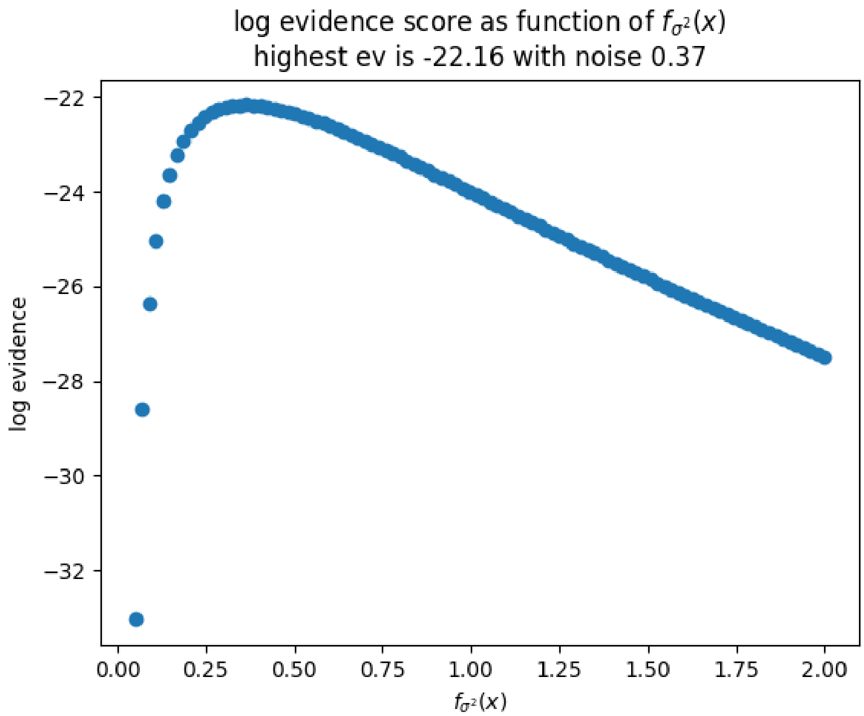

f4 evidence versus degree
0
−100
−200
−300
−400
−500
log evidence
2
3
4
5
6
7
8
9
10
degree


f4 best and worst BLR model fit
True points
worst fit, degree 2
best fit, degree 5
worst confidence
best confidence
$f_4(x)$
x
−3
−2
−1
0
1
2
3
4
5
6
7


f5 evidence versus degree
−600
−800
−1000
−1200
−1400
−1600
−1800
−2000
log evidence
2
3
4
5
6
7
8
9
10
degree


f5 best and worst BLR model fit
2
0
−2
−4
−6
−8
$f_5(x)$
−3
−2
−1
0
1
2
3
x
True points
worst fit, degree 2
best fit, degree 7
worst confidence
best confidence

חלק 2.2



שאלה 7

ניתן לראות שקיבלנו שה-$log\_evidence$ הגבוהה ביותר מתקבל עבור רעש 0.37 למרות שידוע כי הרעש האמיתי הוא 0.25.

לכן נסיק שהערך בו מתקבל המקסימום הוא לא בהכרח ערך הרעש האמיתי.