1.1) a) $P(Y_i = y_i \mid x_{i1}, x_{i2}; c) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - c_1 x_{i1} - c_2 x_{i2})^2}{2\sigma^2}\right)$

b) $L(\{y_i\} \mid \{x_i\}, c) = \log \prod_{i=1}^{n} P(Y_i = y_i \mid x_{i1}, x_{i2}; c) = -\sum_{i=1}^{n} \frac{(y_i - x_i^T c)^2}{2\sigma^2} - n\ln\sqrt{2\pi}\sigma$

c) $$\frac{\partial L}{\partial c} = \sum_{i=1}^{n} \frac{(y_i - x_i^T c) x_i}{\sigma^2} = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i = \left(\sum_{i=1}^{n} x_i x_i^T\right) c$$

定义 $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$ 为数据采样方列的内积定义（易证明该定义满足内积）

$\langle x_1, y \rangle = \sum_{i=1}^{n} x_{i1} y_i$

$\langle x_1, x_2 \rangle = \sum_{i=1}^{n} x_{i1} x_{i2}$

有 $$\begin{bmatrix} \langle y, x_1 \rangle \\ \langle y, x_2 \rangle \end{bmatrix} = \begin{bmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle \\ \langle x_1, x_2 \rangle & \langle x_2, x_2 \rangle \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

有 $$c_1 = \frac{\langle x_2, x_2 \rangle \langle y, x_1 \rangle - \langle x_1, x_2 \rangle \langle y, x_2 \rangle}{\langle x_1, x_1 \rangle \langle x_2, x_2 \rangle - \langle x_1, x_2 \rangle^2}$$

根据内积的柯西不等式知 $\langle x_1, x_1 \rangle \langle x_2, x_2 \rangle \geq \langle x_1, x_2 \rangle^2$

且只在每一组数据的 $x_{i2}/x_{i1} = \lambda = const$ 时取到等号（认为数据不可能满足这种情形）

$$c_2 = \frac{\langle x_1, x_1 \rangle \langle y, x_2 \rangle - \langle x_1, x_2 \rangle \langle y, x_1 \rangle}{\langle x_1, x_1 \rangle \langle x_2, x_2 \rangle - \langle x_1, x_2 \rangle^2}$$

1.2) a) $P(Y_i = y_i \mid x_{i1}, x_{i2}; c) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(y_i - c_1 x_{i1} - c_2 x_{i2})^2}{2\sigma_i^2}\right)$

b) $L = -\sum_{i=1}^{n} \frac{(y_i - x_i^T c)^2}{2\sigma_i^2} - \sum_{i=1}^{n} \ln\sigma_i$ ~~+ const~~

c) $$\frac{\partial L}{\partial c} = \sum_{i=1}^{n} \frac{x_i (y_i - x_i^T c)}{\sigma_i^2} = 0 \Rightarrow \sum_{i=1}^{n} \frac{x_i}{\sigma_i} \frac{y_i}{\sigma_i} = \left(\sum_{i=1}^{n} \left(\frac{x_i}{\sigma_i}\right)\left(\frac{x_i}{\sigma_i}\right)^T\right) c$$

再利用上题中内积定义 即 $\langle \frac{a}{\sigma}, \frac{b}{\sigma} \rangle_n = \sum_{i=1}^{n} \frac{a_i}{\sigma_i} \frac{b_i}{\sigma_i}$

有 $$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \frac{1}{\langle \frac{x_1}{\sigma}, \frac{x_1}{\sigma} \rangle \langle \frac{x_2}{\sigma}, \frac{x_2}{\sigma} \rangle - \langle \frac{x_1}{\sigma}, \frac{x_2}{\sigma} \rangle^2} \begin{bmatrix} \langle \frac{x_2}{\sigma}, \frac{x_2}{\sigma} \rangle \langle \frac{y}{\sigma}, \frac{x_1}{\sigma} \rangle - \langle \frac{x_1}{\sigma}, \frac{x_2}{\sigma} \rangle \langle \frac{y}{\sigma}, \frac{x_2}{\sigma} \rangle \\ \langle \frac{x_1}{\sigma}, \frac{x_1}{\sigma} \rangle \langle \frac{y}{\sigma}, \frac{x_2}{\sigma} \rangle - \langle \frac{x_1}{\sigma}, \frac{x_2}{\sigma} \rangle \langle \frac{y}{\sigma}, \frac{x_1}{\sigma} \rangle \end{bmatrix}$$

Note: 相当于把所有 sample 中的 $x_i \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix}$ 和 y_i 缩小 σ_i 倍并作 1.1 中的估计。

a) $P(Y_i = y_i \mid x_i; b) = \frac{1}{2b} \exp\left(-\frac{|y_i - c^T x_i|}{b}\right)$

b) $L(\{y_i\} \mid \{x_i\}; b) = -\sum_{i=1}^{n} \frac{|y_i - c^T x_i|}{b} - n \ln(2b)$

c) 高斯型噪声假设会给远处的点(误差大的点)过多的注意，若数据中有极少数个离群点 outliers 与其它数据差很远，高斯型噪声模型会导致我们的参数估计被这些 outliers 影响较大，而 Laplace 噪声模型相对来说对噪声或者是 outlier 对其造成的影响没有对 Gaussian Error Model 大。

a) ~~若 sample 在 t 上~~ 对于 $0 \le x \le a$，对 $p_n(x)$ 的估计是由采样过程中 $0 \le t_i \le x$ 的样本点加权得到的，即

$$p_n(x) = \frac{1}{n h_n} \sum_{i=1}^{n} e^{-\frac{(x-t_i)}{h_n}} I(t_i \le x)$$

$$E(p_n(x)) = \frac{1}{n h_n} \sum_{i=1}^{n} E_{t \sim p}\left[I(t_i \le x)\, e^{-\frac{(x-t_i)}{h_n}}\right]$$

由于 t_i 为 i.i.d sample

$$\therefore E(p_n(x)) = \frac{1}{h_n} E_{t\sim p}\left[I(t \le x)\, e^{-\frac{(x-t)}{h_n}}\right]$$

$$= \frac{1}{h_n} \int_0^x p(t)\, e^{\frac{t-x}{h_n}}\, dt$$

$$\overset{z = \frac{t-x}{h_n}}{=} \int_{-\frac{x}{h_n}}^{0} \frac{1}{a} e^{z}\, dz$$

$$= \frac{1}{a}\left(1 - e^{-\frac{x}{h_n}}\right)$$

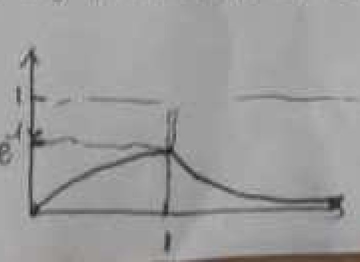
对于 $a \le x$ 的点 x 有

$$E(p_n(x)) = \frac{1}{h_n}\int_0^a p(t)\, e^{\frac{t-x}{h_n}}\, dt = \frac{1}{a}\int_{-\frac{x}{h_n}}^{\frac{a-x}{h_n}} e^{\frac{t-x}{h_n}}\, d\frac{t-x}{h_n}$$

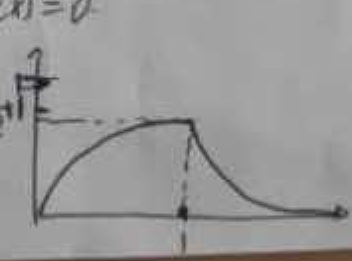
$$= \frac{1}{a} e^{-\frac{x}{h_n}}\left(e^{\frac{a}{h_n}} - 1\right)$$

对于 $x < 0$，由于 $P(t_i \le x) = 0$，$\therefore \bar{p}_n(x) = 0$

b) $h_n = 1$



$h_n = \frac{1}{4}$



$h_n = \frac{1}{16}$

c) 在编程适于0的x处，$b(x)$最大，$\left(\frac{1}{a} - \frac{1}{a}\left(1 - e^{-\frac{x}{h_n}}\right)\right) = 0.01 \cdot \frac{1}{a}$

其中x取 0.01a

$\Rightarrow$ $e^{-\frac{0.01a}{h_n}} = 0.01 \Rightarrow h_n = 0.00217a$

d) $h_n = 0.00217$

此图见pdf文档

3\. a) 在题中假设情况下，判断错误只可能是：来表在n个sample中的 ~~数量~~ 数量 ~~该~~ $j \le \frac{k-1}{2}$

即 $P_n(e) = \sum_{j=0}^{\frac{k-1}{2}} \binom{n}{j} \left(\frac{1}{2}\right)^j \left(\frac{1}{2}\right)^{n-j} \cdot \frac{1}{2} + \sum_{j=0}^{\frac{k-1}{2}} \binom{n}{j} \left(\frac{1}{2}\right)^j \left(\frac{1}{2}\right)^{n-j} \frac{1}{2}$

$= \left(\frac{1}{2}\right)^n \sum_{j=0}^{\frac{k-1}{2}} \binom{n}{j}$

b) $P_n(e)$ 是一个随 k 单调递增（因为 k 满足 $\le n$）

$\therefore$ $k=1$ 时有比 $k>1$ 更小的 error rate

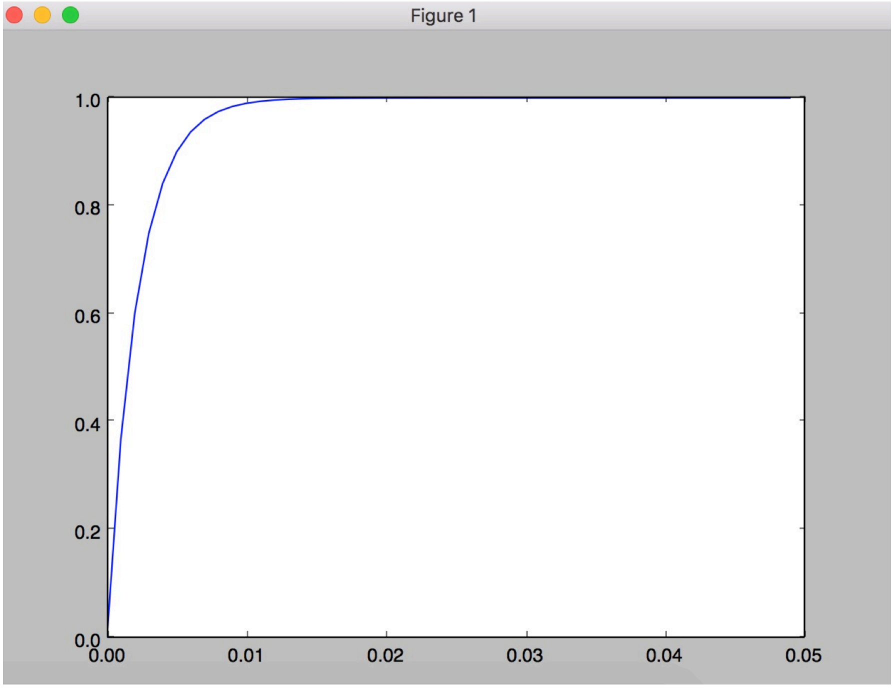
c) $P_n(e) = \frac{1}{2^n} \sum_{j=0}^{\frac{k-1}{2}} \binom{n}{j} < \frac{1}{2^n} \frac{a\sqrt{n}-1}{2} \binom{n}{\frac{a\sqrt{n}-1}{2}} = \frac{1}{2^n} \frac{1}{n - \frac{a\sqrt{n}-1}{2}} \binom{n}{\frac{a\sqrt{n}-3}{2}}$

由于 $\frac{1}{2^n} \binom{n}{\frac{a\sqrt{n}-3}{2}} < 1$ $\therefore$ $P_n(e) < \frac{1}{n - \frac{a\sqrt{n}-1}{2}}$

$\therefore P_n(e) \xrightarrow{n \to \infty} 0$

2.

d) 0~0.05 中的概率密度函数如图所示：



4.

a)

代码实现见附件：

入口脚本为 parzen.m。

gen_sample.m 生成题目中的数据。

gaussian_window.m 和 rect_window.m 两个文件实现用不同的 parzen 窗对概率密度函数进行估计。

eval_error.m 高阶函数实现多次采样，并用不同参数调用不同 parzen 窗函数，并返回概率密度函数的平方错误。

b)

这个概率分布函数的平方误差可用 Riemann 积分公式写为:

$$\int [p_n(x) - p(x)]^2 dx = \sum_i [p_n(x_i) - p(x_i)]^2 \delta x$$

其中$x_{i+1} - x_i = \delta x$为在 x 坐标轴上的均匀采点。

c)

固定 n = 10000:

在不同 a 的取值下，通过 10 次不同的 n 个 sample 的采样进行的非参数估计得到的错误的均值和方差如下：

a	二次错误 mean	二次错误 variance
10	0.1074	0.0008 e-5
1	0.0002	0.0008 e-5

0.1	0.0009	0.0006 e-5
0.01	0.01	0.0159 e-5
0.001	0.0995	0.5174 e-5

从画出来的图和错误都能大致看出在 n 取 10000 时，a 取 0.1 ~ 1 相比其他取值较好，有着较小的 squared error mean(更准确的估计)和 square error variance(更稳定、一致的估计)。

下面是固定 a = 0.1，使用不同采样数量 n，得到的 square error mean 和 squared error variance。可以看出在 a 固定的情况下(且不是十分大)，如果采样点越多，错误均值越小，且错误方差也减小，这符合直观。二次错误的均值基本遵循：增加十倍采样，降低 10 倍的规律。

n	二次错误 mean	二次错误 variance
10	0.9276	4.250 e-2
100	0.0981	1.789 e-4
1000	0.0096	2.507 e-6
10000	0.0009	1.513 e-8
100000	0.0001	4.425 e-10

d)

选择$a = \frac{10000}{n}$ 左右，如果选取太大的 a，会导致分辨率太低，所有的采样点对大多 x 的贡献都一致；如果选取太小的 a，虽然提高了分辨率，但是在 n 不够大时，不少点周围没有足够多的样本，会导致非参数估计得到的概率分布十分不光滑。

e)
高斯窗：
固定 n = 10000:
在不同 sigma 的取值下，通过 5 次不同的 n 个 sample 的采样进行非参数估计得到的错误的均值和方差如下：
由于高斯窗函数做估计的函数比矩形窗估计运行慢（已用近似 support 集优化，仍需多次计算 normpdf)。所以只跑了 5 次。

sigma	二次错误 mean	二次错误 variance
0.5	0.0012	3.7549e-8
0.1	3.05e-4	3.8922e-9
0.01	0.0030	1.1412e-7
0.001	0.0272	2.4387e-6

从上表结果可以看出 n=10000 时，sigma 取 0.1，错误的均值和方差较小。

固定 sigma = 0.1，使用不同的采样数量 n，每个不同的采样次数跑 5 次，得到的错误的均值和方差如下：

n	二次错误 mean	二次错误 variance

10	0.2252	5.60e-3
100	0.0228	3.57e-5
1000	0.0024	1.57e-7
10000	2.744e-4	4.33e-9
100000	2.694e-5	7.31e-11

在高斯窗实现过程中，为了防止太慢，对每个 sigma 计算了一个，按照 norminv(1-1e-3,0, sigma)作为半边的 support 长度，忽略 support_threshold 以外的所有点，得到了极大的速度提升。

f)

可以大概由高斯分布的 3sigma 定律知道应该也和矩形窗的 a 取值方法类似，可以取 sigma 为 n 的一个反比例函数，从实验中估计大概为$sigma = \frac{1000}{n}$。可以注意到如果用我们的经验公式取 a 和 sigma，用高斯窗函数估计概率对于每个 x 需要计算的有影响点比矩形窗要少，因为如果以 1e-3 作为 support threshold，有 $3sigma < a/2$。每个样本点的支撑范围比矩形窗估计要小，且可以达到类似的错误均值和小很多的错误方差(见两个窗函数评估中的第一个表格里的标红数据)。