Regression: predicting $y$ from $x$, $p(y|x)$

- For simplicity, first assume a linear predictor
  
  $$y = ax + b + \epsilon$$

- Gaussian noise $N(0, \sigma)$

- To be fancy, let $y$ be a linear sum of basis functions
  
  $$y = \sum_i \alpha_i \phi_i(x) + \epsilon$$

- We find $\alpha_i$ by least squares. Let $\hat{\alpha} = [\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_n]^T$
  
  $$\hat{\alpha} = \arg\min_{\alpha} \sum_i \frac{(y_i - \sum_j \alpha_j \phi_j(x_i))^2}{2\sigma^2}$$

  We find set of coefficients that minimizes squared error between actual $y_i$'s and predicted $\hat{y}_i$'s.

Aside: A common problem is that the input features are correlated with each other.

This is often the case when $n$ is large ("curse of dimensionality")

1. Consider the area of the surface of a cube in $n$ dimensions as the difference of the volumes of $n$ cubes, one inside the other:

   $$\text{volume} = 1^n - (-1)^n \to \frac{1}{n} \text{ as } n \to \infty.$$  

   - all volume is located on skin of cube
   - points randomly distributed in cube will tend to lie on surface whose correlations

   - the components of $\hat{x}$ are correlated

   - use a regularizer term in the minimization

   $$\hat{\alpha} = \arg\min_{\alpha} \sum_i \frac{(y_i - \sum_j \alpha_j \phi_j(x_i))^2}{2\sigma^2} + \lambda \sum_j \alpha_j^2$$

   - we penalize large coefficients $\alpha_i$ because they suggest we are fitting to noise/correlations

   - we can set $\lambda$ by cross validation; we leave out one data point $(x_i, y_i)$, fit to the rest, and predict the $i$th point $\Rightarrow$ goodness of prediction $\square$
EM

1. **Motivation**: In regression setup, it'd be nice to throw away outliers & fit to "true" data

   ![Diagram](image)

   Imagine yi's are generated by 2 processes:
   - With prob $\pi$, $y = ax + b + \epsilon$
   - With prob $1-\pi$, $y \sim \text{uniform noise}$

   We can think of each data point $(x_i, y_i)$ as coming with an extra label specifying which process generated it:

   $X_i = (x_i, y_i, \pi_i)$

   **Model parameters**: $a, b, \pi$

1. Assume we have an estimate of our model $a^{(n)}$, $b^{(n)}$, $\pi^{(n)}$.

   We now want to estimate $\pi_i$ by $E[\pi_i | a^{(n)}, b^{(n)}, \pi^{(n)}]$

   

   $E[\pi_i | a^{(n)}, b^{(n)}, \pi^{(n)}] = \frac{P(C_i = 1 | a^{(n)}, b^{(n)}, \pi^{(n)})}{P(C_i = 0 | a^{(n)}, b^{(n)}, \pi^{(n)}) + P(C_i = 1 | a^{(n)}, b^{(n)}, \pi^{(n)})}$

   

   2. Given $E[\pi_i | \ldots] \pi_i$, we can re-estimate $a^{(n+1)}$, $b^{(n+1)}$, $\pi^{(n+1)}$ with extraneous weighted least square:

   $a^{(n+1)}$, $b^{(n+1)} = \arg \min_{a,b} \sum_{i=1}^{N} w_i \frac{(y_i - ax - b)^2}{s_{\pi_i}^2}$

   where $w_i = E[\pi_i | \ldots]$

   We can set $\pi^{(n)} = \frac{1}{N} \sum_{i=1}^{N} w_i$
Qi Shum cont'd

Hidden variable is binary vector with a '1' specifying which model this frame belongs to, equivalent to $T_i$ from before.

For each model, we see in 60-112 frame sequences in the data.

Problems:
1) We don't have enough data to fit the dynamics for each of our 246 models.
2) Can't compute $E$-step of EM.

$$E [S_i | \text{model}] =$$

- We first need to calculate the likelihood of observed frames.
- This term not tractable.

For model (a):

$$P(y_i^{(a)} = \text{measured} | \sigma_i = a_j, \ldots) = \sum_{\sigma_i} P(y_i^{(a)} | \text{measured}) P(X^{(a)}) \mathcal{N}$$

For model (b):

$$P(y_n^{(b)} = \text{measured} | \sigma_n = a_j, \ldots) \approx \text{exponential in } N$$

For each previous frame, we need to consider both cases of model (a) and model (b).