Given data $(X_1, Y_1), \ldots, (X_n, Y_n) \sim F$

dist in high dimension

Goal: Find β minimizing

$$R(\beta) = \mathbb{E}\left[L(r_\beta(X), Y)\right] = \int L(r_\beta(x), y)\, dF$$

where $L(\hat{y}, y)$ is loss function

Since we don't know F, replace w. empirical CDF F_n

$$R_n(\beta) = \int L(r_\beta(x), y)\, dF_n = \frac{1}{n}\sum_{i=1}^{n} L(r_\beta(X_i), Y_i)$$

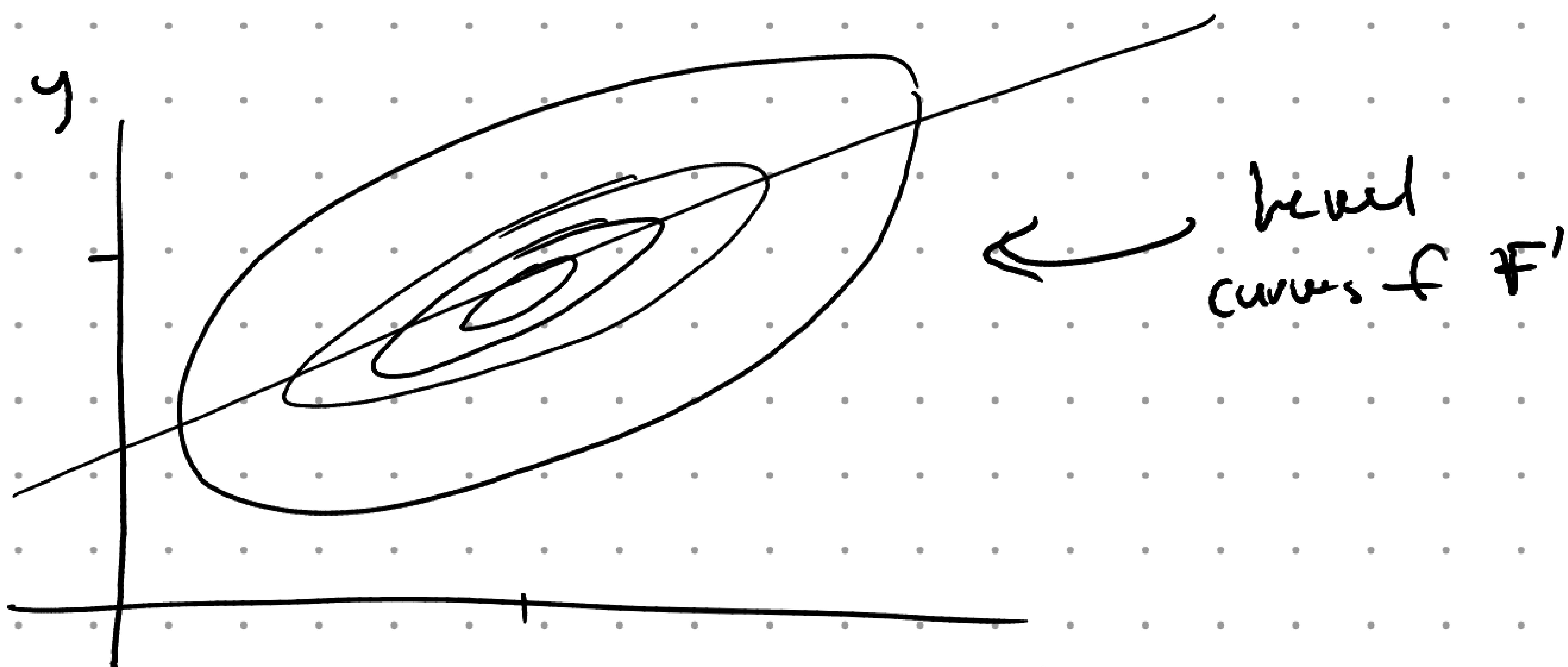
Now minimize $R_n(\beta)$ to find $\hat{\beta}_n$

How good is this estimate? IE what is $R(\hat{\beta}_n)$?

Ex

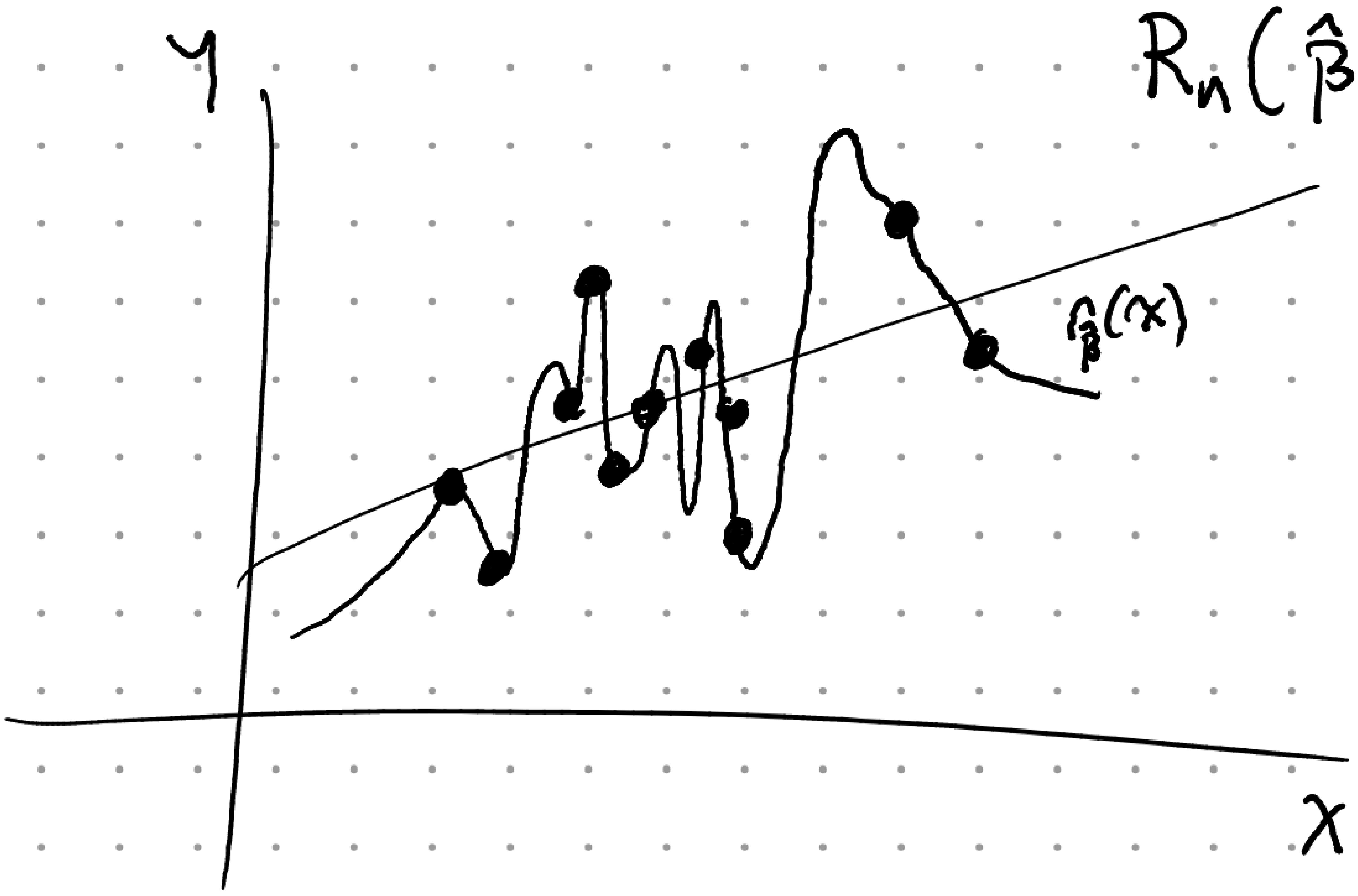
Suppose X, Y joint gaussian

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY}^2 \\ \sigma_{XY}^2 & \sigma_Y^2 \end{bmatrix}\right)$$



$$r(X) = \mathbb{E}[Y|X] = \min_r \mathbb{E}\left[(r(X) - Y)^2 / X\right]$$

$$R_n(\hat{\beta}) = 0$$



Intuitively, our model was <u>too complex</u> so we <u>overfit</u> the data. If we restricted to a simpler model, then we would be less likely to overfit. But simpler models may not be expressive enough to represent trends in the data

Here, $R_n(\hat{\beta}) < R(\beta)$

This is not uncommon outcome.

To estimate $R(\beta)$, we can use new data $(\tilde{X}_1, \tilde{X}_2), \dots, (\tilde{X}_{\tilde{n}}, \tilde{Y}_{\tilde{n}})$ and compute $\tilde{R}_{\tilde{n}}(\hat{\beta})$

New data often not available, so split original data into two sets, one for learning β, one for evaluating the quality.

Least Squares as MLE

Suppose $X_1, \ldots, X_n$ fixed set of points
and $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$$\varepsilon_i \sim N(0, \sigma^2) \quad \text{iid}$$

Then $Y_i \sim N(\mu_i, \sigma^2)$, $\mu_i = \beta_0 + \beta_1 X_i$

$$\mathcal{L}_n(\beta) \propto \prod_{i=1}^{n} f_\beta(Y_i)$$

density for Y_i

$$\propto \prod_{i=1}^{n} \exp\left(-\frac{(Y_i - \mu_i)^2}{2\sigma^2}\right)$$

$$= \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mu_i)^2\right)$$

$$\max_\beta \mathcal{L}_n(\beta) \Leftrightarrow \min_\beta \sum_{i=1}^{n} (Y_i - \mu_i)^2$$

$$= \min_\beta \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$