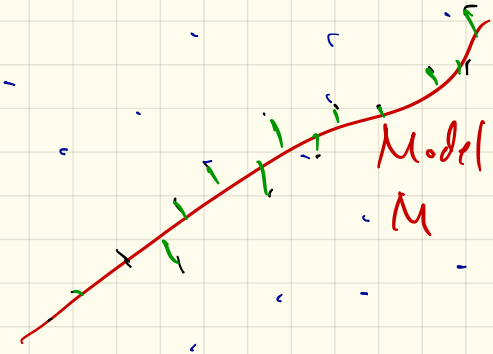
Noise in Data

- Spurious Readings
  outliers



- Measurement Error

Gaussian $\approx$ sum of squares error $\|\ \|_2$

- Background Data
  - missing data

- Adversarial Data

Cross-Validation + Regularization

Model $M(x) = \underset{M}{\operatorname{argmin}} \; \| M(x) - x \|_2^2 + s \|M\|_2^2$

lasso

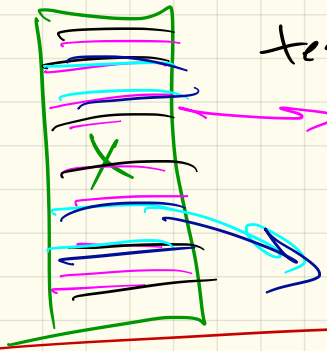
- biases towards simpler models

observed data

priors

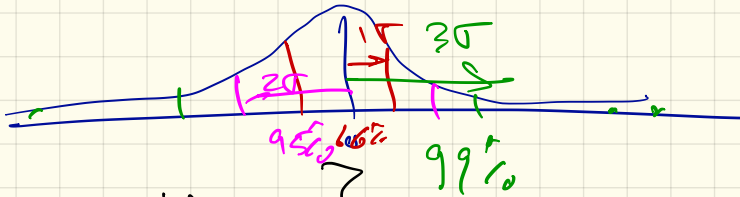
regularization parameter

choose s w/ C-V



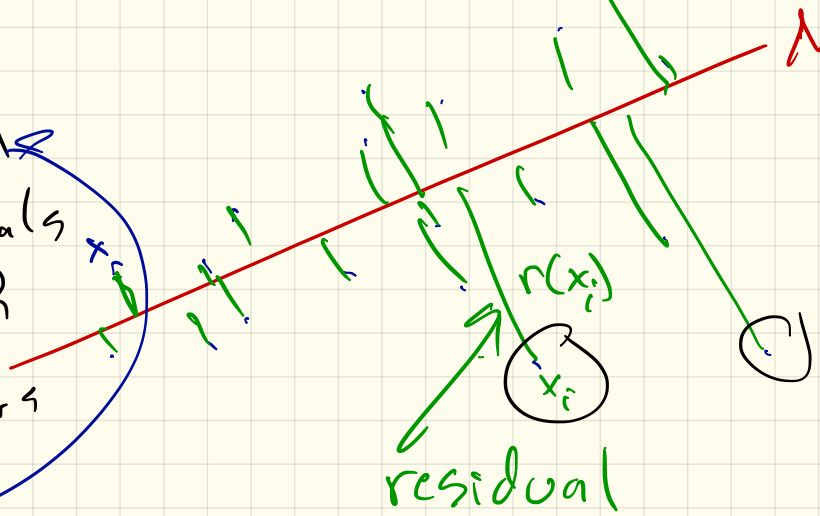
Bootstrapping | for $i=1$ to b
draw n points $X_i \leftarrow X$ with replacement
measure model $M(x_1), M(x_2) \ldots M(x_b)$

Outliers



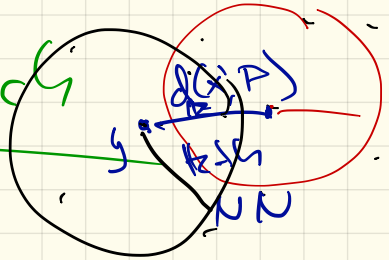
- ~~How to~~ Should I remove them?

1. Build Model M
2. Calculate residuals
   $\{r(x) \mid x \in X\}$
3. Remove outliers
   $r(x) > \tau$
4. Go to 1

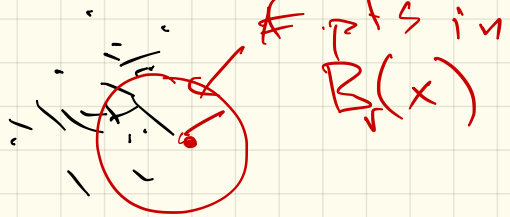


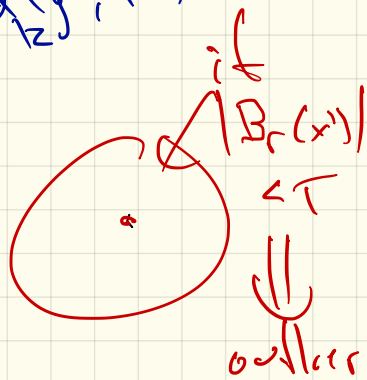
outliers $= x \in X$ w/ largest residual $r(x_i)$

Density-based Approach



$d_k(y, P)$

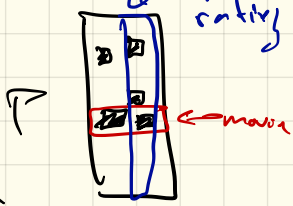




reverse-nearest neighbor

Missing Data

Data set $P \in \mathbb{R}^{n \times d}$ matrix



$$\Omega = \{(i,j) \in [n] \times [d] \mid P_{ij} \neq \phi\}$$

$\Pi_\Omega(P)$ ← elements w/ score

$\Pi_\Omega^{\perp}(P)$ ← elements w/ out scores

$$P^* = \underset{X}{\operatorname{argmin}} \; \|\Pi_\Omega(P - X)\|_F^2 + \lambda \|X\|_*$$

nuclear norm = sum sing. values

Matrix completion (P, Ω, λ)

0. Initialize X : $X_{ij} \leftarrow$ average of observed entries in row/column

or $\Pi_{\Omega}(P)$

repeat

1. $U S V^T \leftarrow \text{svd}(X)$
2. $\hat{X} \leftarrow U \boxed{\phi_\lambda(S)} V^T$
3. $X \leftarrow \Pi_{\Omega}(P) + \Pi^{\perp}_{\Omega}(\hat{X})$

until ("converges")

Return $\hat{X}$

$\phi_\lambda(S) = \text{diag}\big((S_{11}-\lambda)_+, (S_{22}-\lambda)_+, \ldots, (S_{dd}-\lambda)_+\big)$

$(x-\lambda)_+ = \max\{0, x-\lambda\}$