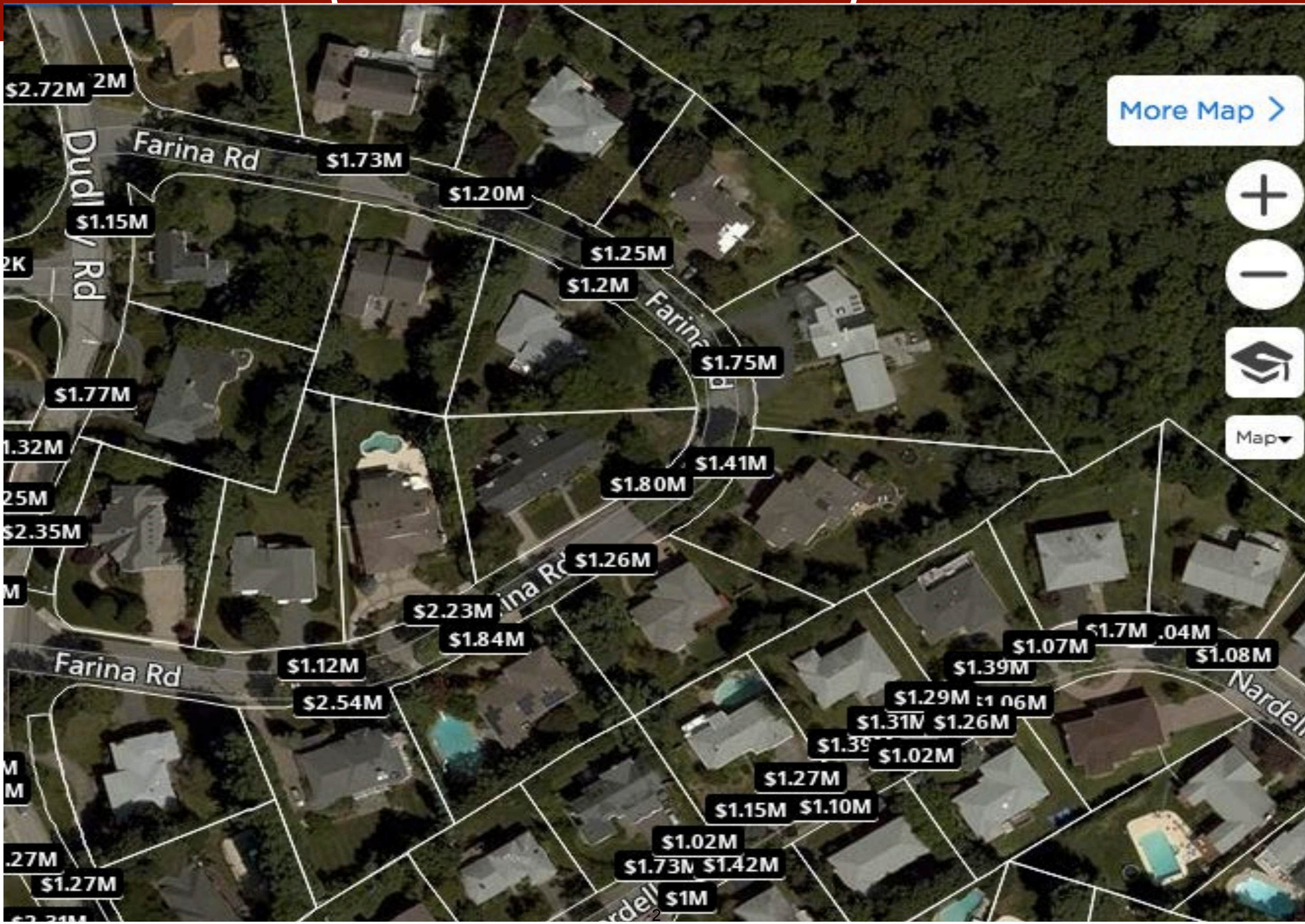
Normal equations
Random processes

6.011, Spring 2018

Lec 15

Zillow (founded 2006)



Zestimates

How do we come up with the Zestimate and what's in the formula?

We use proprietary automated valuation models that apply advanced algorithms to analyze our data to identify relationships within a specific geographic area, between this home-related data and actual sales prices. Home characteristics, such as square footage, location or the number of bathrooms, are given different weights according to their influence on home sale prices in each specific geography over a specific period of time, resulting in a set of valuation rules, or models that are applied to generate each home's Zestimate. Specifically, some of the data we use in this algorithm include:

Physical attributes: Location, lot size, square footage, number of bedrooms and bathrooms and many other details.

Tax assessments: Property tax information, actual property taxes paid, exceptions to tax assessments and other information provided in the tax assessors' records.

Prior and current transactions: Actual sale prices over time of the home itself and comparable recent sales of nearby homes

Currently, we have data on 110 million homes and Zestimates and Rent Zestimates on approximately 100 million U.S. homes. (Source: Zillow Internal, March 2013)

LMMSE for multivariate case

$$\min_{a_0,\ldots,a_L} E[(Y - \underbrace{\{a_0 + \Sigma_{j=1}^{L} a_j X_j\}}_{\widehat{Y}_\ell})^2]$$

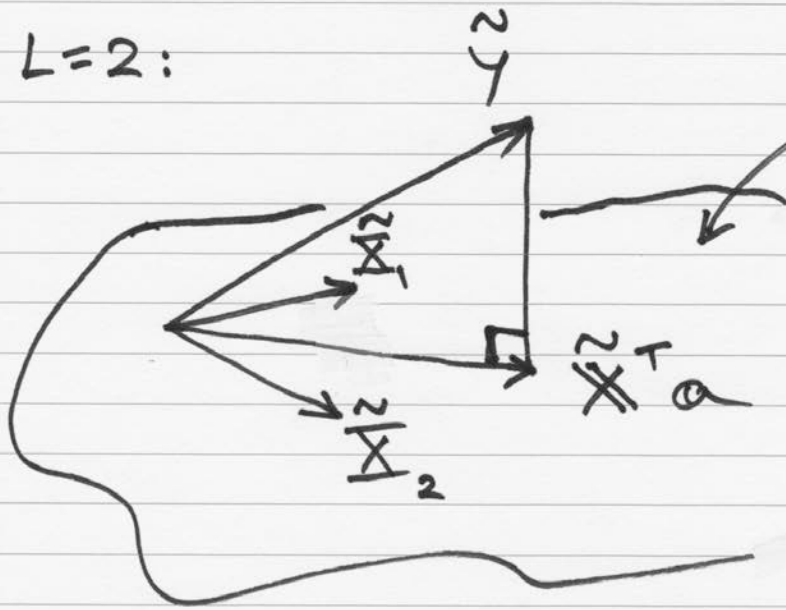
$$\text{First } \min_{a_0} \Rightarrow a_0 = \mu_Y - \Sigma_{j=1}^{L} a_j \mu_{X_j}$$

This ensures unbiasedness of the estimator.

$$\text{Now } \min_{a_1,\ldots,a_L} E[(\widetilde{Y} - \Sigma_{j=1}^{L} a_j \widetilde{X}_j)^2]$$

Geometric picture

$L=2$:



Plane spanned by $\tilde{X}_1, \tilde{X}_2$

$$\Rightarrow \tilde{Y} - \tilde{X}^T a \perp \tilde{X}_j$$

i.e.,

$$E[\tilde{X}(\tilde{Y} - \tilde{X}^T a)] = 0$$

$$\Rightarrow C_{XX}\, a = c_{XY}$$ NORMAL EQUATIONS

$$\left.\begin{aligned} \underbrace{\tilde{Y} - \tilde{\hat{Y}}_{lin}} &\perp \tilde{X}_j \; (= X_j - \mu_{X_j}) \\ Y - \hat{Y}_{lin} &\perp X_j \end{aligned}\right\}$$ ORTHOGONALITY RELATIONS

Applying orthogonality gives the "normal equations"

$$E\left[\left(\widetilde{Y} - \Sigma_{j=1}^{L}\, a_j \widetilde{X}_j\right)\widetilde{X}_i\right] = 0$$

$$\begin{bmatrix} \sigma_{X_1 X_1} & \sigma_{X_1 X_2} & \cdots & \sigma_{X_1 X_L} \\ \sigma_{X_2 X_1} & \sigma_{X_2 X_2} & \cdots & \sigma_{X_2 X_L} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_L X_1} & \sigma_{X_L X_2} & \cdots & \sigma_{X_L X_L} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_L \end{bmatrix} = \begin{bmatrix} \sigma_{X_1 Y} \\ \sigma_{X_2 Y} \\ \vdots \\ \sigma_{X_L Y} \end{bmatrix}$$

$$\left(\mathbf{C}_{\mathbf{XX}}\right)\mathbf{a} = \mathbf{c}_{\mathbf{X}Y}$$

$$\text{MMSE:} \quad \sigma_Y^2 - \mathbf{c}_{Y\mathbf{X}}\left(\mathbf{C}_{\mathbf{XX}}\right)^{-1}\mathbf{c}_{\mathbf{X}Y} = \sigma_Y^2 - \mathbf{c}_{Y\mathbf{X}}.\mathbf{a}$$

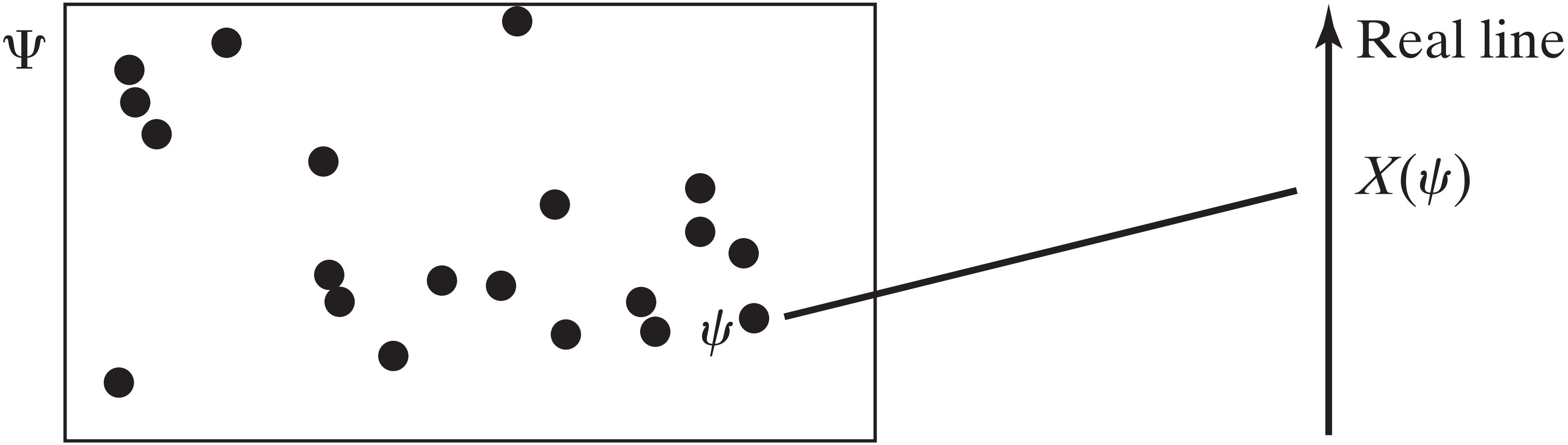
Estimating mean vector and covariance matrix from data

Given N independent measurements: $\mathbf{X}_i \;, \quad i = 1, \cdots, N$

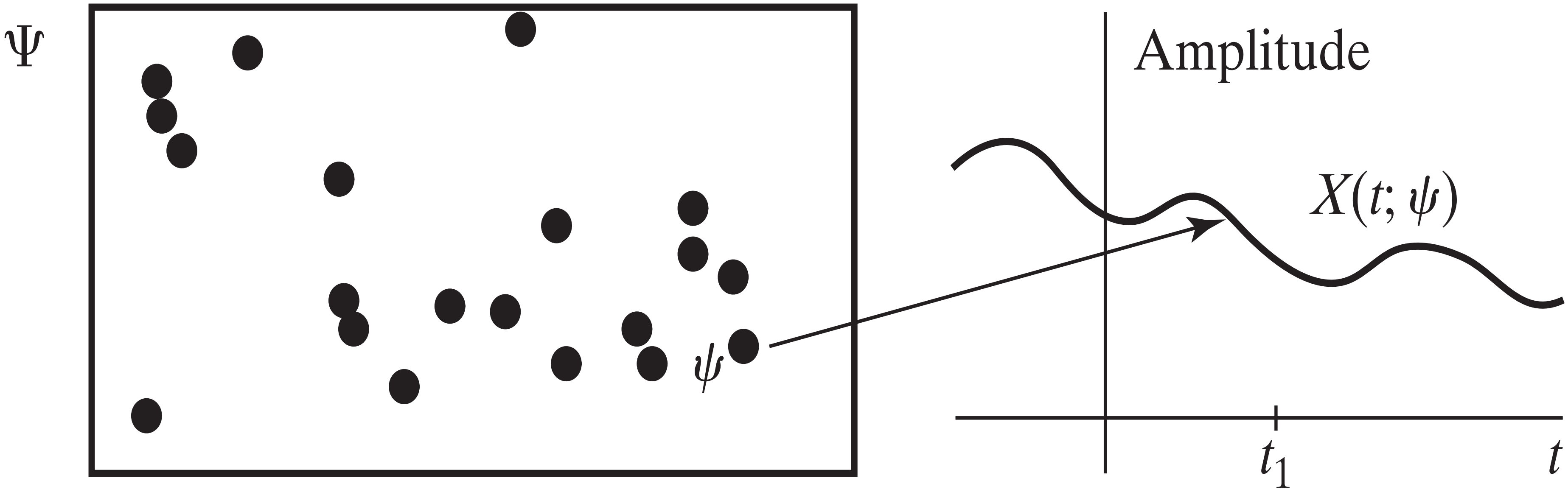
Estimate of mean: $\widehat{\mu}_{\mathbf{X}} = \frac{1}{N} \sum_{1}^{N} \mathbf{X}_i$

Estimate of covariance: $\widehat{\mathbf{C}}_{\mathbf{X}\mathbf{X}} = \frac{1}{N-1} \sum_{1}^{N} (\mathbf{X}_i - \widehat{\mu}_{\mathbf{X}})(\mathbf{X}_i - \widehat{\mu}_{\mathbf{X}})^T$

Random variable



Random process



Signal ensemble for outcomes a,b,c,d; & determination of $R_{XX}(t_1,t_2)$