Linear Regression: Linear Minimum MSE Estimation

Probability and Statistics for Data Science

Carlos Fernandez-Granda

NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

NYU DATA SCIENCE

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at `https://www.ps4ds.net`

Regression

Goal: Estimate response from features

Previously: Estimate response from **one** feature

This video: Estimate response from **multiple** features

Probabilistic formulation

Find function h, such that $h(x)$ approximates the response $\tilde{y}$ when the features $\tilde{x} = x$

How do we evaluate the estimator?

Mean squared error (MSE): $\mathrm{E}\left[(\tilde{y} - h(\tilde{x}))^2\right]$

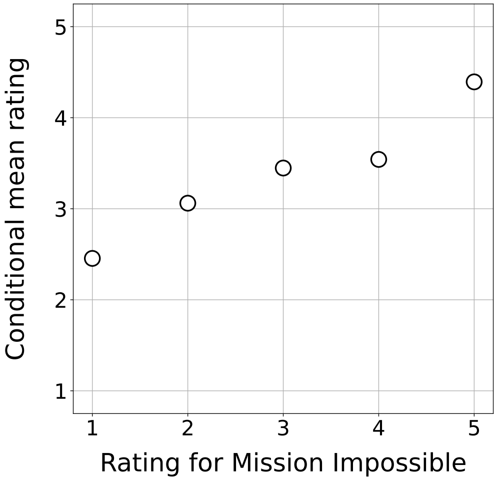
MMSE estimator

The conditional mean is the minimum MSE estimator

$$\mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x}) = \arg\min_{h(\tilde{x})} \mathrm{E}\left[(\tilde{y} - h(\tilde{x}))^2\right]$$

Movie rating

Estimate rating for Independence Day given rating for Mission Impossible



Are we done here? No!

Computing conditional mean is often impossible due to curse of dimensionality

To predict from 100 movie ratings, how many conditional averages do we need to estimate?

$5^{100} > 10^{68}$!

Linear regression

We approximate the response as an affine function of the features

$$\begin{aligned}\tilde{y} \approx \ell(\tilde{x}) &:= \sum_{i=1}^{d} \beta[i]\tilde{x}[i] + \alpha \\ &= \beta^T \tilde{x} + \alpha\end{aligned}$$

Linear minimum MSE (MMSE) estimator

$$(\beta_{\text{MMSE}}, \alpha_{\text{MMSE}}) = \arg\min_{\beta,\alpha} \mathrm{E}\left[\left(\tilde{y} - \beta^T \tilde{x} - \alpha\right)^2\right]$$

$$\ell_{\text{MMSE}}(\tilde{x}) := \beta_{\text{MMSE}}^T \tilde{x} + \alpha_{\text{MMSE}}$$

One feature: $\tilde{a}$

$$
\begin{aligned}
\ell_{\text{MMSE}}(a) &:= \beta_{\text{MMSE}} a + \alpha_{\text{MMSE}} \\
&= \sigma_{\tilde{y}} \, \rho_{\tilde{a},\tilde{y}} \left(\frac{a - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}} \right) + \mu_{\tilde{y}}
\end{aligned}
$$

$\mu_{\tilde{a}}$, $\mu_{\tilde{y}}$: means of $\tilde{a}$ and $\tilde{y}$

$\sigma_{\tilde{a}}$, $\sigma_{\tilde{y}}$: standard deviations of $\tilde{a}$ and $\tilde{y}$

$\rho_{\tilde{a},\tilde{y}}$: correlation coefficient of $\tilde{a}$ and $\tilde{y}$

Strategy to derive linear minimum MSE estimator

- Derive best additive constant $\alpha^*(\beta)$ as a function of linear coefficients
- Plug in $\alpha^*(\beta)$ to express MSE as a function of β
- Minimize MSE to find best linear coefficients

Minimum MSE constant estimate

Best constant estimate of $\tilde{b}$?

$$\arg\min_{c\in\mathbb{R}} \mathrm{E}\left[(c-\tilde{b})^2\right] = \mathrm{E}[\tilde{b}]$$

Additive constant

$$
\begin{aligned}
\alpha^*(\beta) &:= \arg\min_{\alpha} \mathrm{E}\left[(\tilde{y} - \beta^T \tilde{x} - \alpha)^2\right] \\
&= \mathrm{E}\left[\tilde{y} - \beta^T \tilde{x}\right] \\
&= \mathrm{E}\left[\tilde{y}\right] - \beta^T \mathrm{E}\left[\tilde{x}\right] \\
&= \mu_{\tilde{y}} - \beta^T \mu_{\tilde{x}}
\end{aligned}
$$

Linear coefficients

For any β and any α, $\mathsf{MSE}(\beta, \alpha) \geq \mathsf{MSE}(\beta, \alpha^*(\beta))$

$$\beta_{\mathsf{MMSE}} = \arg\min_{\beta} \mathsf{MSE}(\beta, \alpha^*(\beta))$$

Mean squared error

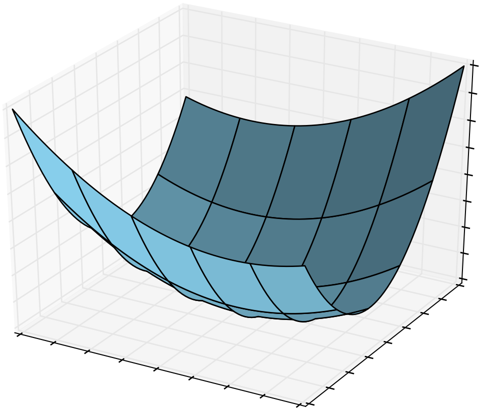
$$\begin{aligned}
&\text{MSE}(\beta, \alpha^*(\beta)) \\
&= \mathrm{E}\left[(\tilde{y} - \beta^T \tilde{x} - \alpha^*(\beta))^2\right] = \mathrm{E}\left[(\tilde{y} - \beta^T \tilde{x} - (\mu_{\tilde{y}} - \beta^T \mu_{\tilde{x}}))^2\right] \\
&= \mathrm{E}\left[(\text{ct}\left(\tilde{y}\right) - \beta^T \text{ct}\left(\tilde{x}\right))^2\right] \\
&= \mathrm{E}\left[\text{ct}\left(\tilde{y}\right)^2\right] + \beta^T \mathrm{E}\left[\text{ct}\left(\tilde{x}\right)\text{ct}\left(\tilde{x}\right)^T\right]\beta - 2\beta^T \mathrm{E}\left[\text{ct}\left(\tilde{x}\right)\text{ct}\left(\tilde{y}\right)\right] \\
&= \sigma_{\tilde{y}}^2 + \beta^T \Sigma_{\tilde{x}} \beta - 2\beta^T \Sigma_{\tilde{x}\tilde{y}} = q(\beta)
\end{aligned}$$

$$\begin{aligned}
\text{ct}\left(\tilde{x}\right) &:= \tilde{x} - \mu_{\tilde{x}} \\
\text{ct}\left(\tilde{y}\right) &:= \tilde{y} - \mu_{\tilde{y}}
\end{aligned}$$

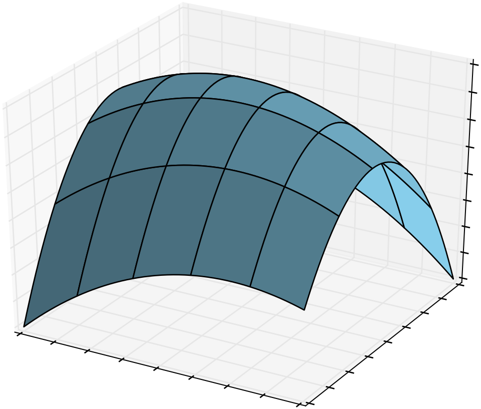
Cross-covariance

$$\Sigma_{\tilde{x}\tilde{y}} := \mathrm{E}\left[\mathrm{ct}\left(\tilde{x}\right)\mathrm{ct}\left(\tilde{y}\right)\right] = \begin{bmatrix} \mathrm{Cov}\left[\tilde{x}[1], \tilde{y}\right] \\ \mathrm{Cov}\left[\tilde{x}[2], \tilde{y}\right] \\ \cdots \\ \mathrm{Cov}\left[\tilde{x}[d], \tilde{y}\right] \end{bmatrix}$$

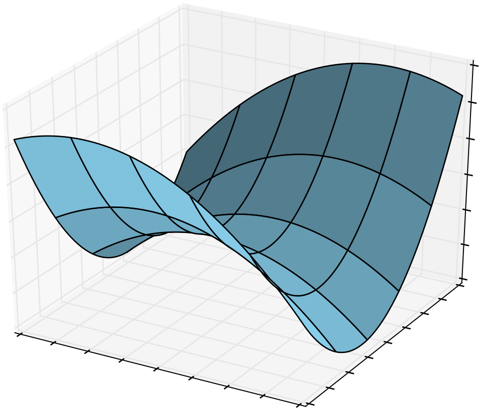
Is $q(\beta)$ convex?



Or concave?



Or neither?



Gradient and Hessian

$$q(\beta) := \sigma_{\tilde{y}}^2 + \beta^T \Sigma_{\tilde{x}} \beta - 2\beta^T \Sigma_{\tilde{x}\tilde{y}}$$

$$\nabla q(\beta) = 2\Sigma_{\tilde{x}} \beta - 2\Sigma_{\tilde{x}\tilde{y}}$$

$$\nabla^2 q(\beta) = 2\Sigma_{\tilde{x}}$$

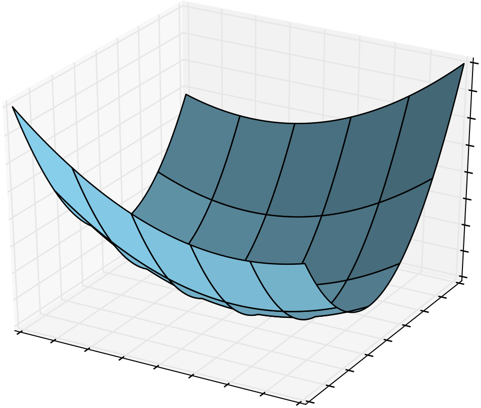
Covariance matrices are positive semidefinite

For any vector $a \in \mathbb{R}^p$

$$a^T \Sigma_{\tilde{x}} a = \mathrm{Var}\left[a^T \tilde{x}\right] \geq 0$$

If $\Sigma_{\tilde{x}}$ is full rank, then positive definite

Convex!



Linear MMSE estimator

$$\nabla q(\beta_{\text{MMSE}}) = 0$$

$$2\Sigma_{\tilde{x}}\beta - 2\Sigma_{\tilde{x}\tilde{y}} = 0$$

$$\beta_{\text{MMSE}} = \Sigma_{\tilde{x}}^{-1}\Sigma_{\tilde{x}\tilde{y}}$$

$$\begin{aligned} \alpha_{\text{MMSE}} &= \alpha^*(\beta_{\text{MMSE}}) \\ &= \mu_{\tilde{y}} - \beta_{\text{MMSE}}^T \mu_{\tilde{x}} \end{aligned}$$

$$\begin{aligned} \ell_{\text{MMSE}}(\tilde{x}) &:= \beta_{\text{MMSE}}^T \tilde{x} + \alpha_{\text{MMSE}} \\ &= \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} (\tilde{x} - \mu_{\tilde{x}}) + \mu_{\tilde{y}} \end{aligned}$$

One feature: $\tilde{a}$

$$
\begin{aligned}
\ell_{\text{MMSE}}(a) &:= \beta_{\text{MMSE}} a + \alpha_{\text{MMSE}} \\
&= \Sigma_{\tilde{a}\tilde{y}}^{T} \Sigma_{\tilde{a}}^{-1} \left(\tilde{a} - \mu_{\tilde{a}} \right) + \mu_{\tilde{y}} \\
&= \frac{\operatorname{Cov}\left[\tilde{a}, \tilde{y}\right]}{\sigma_{a}} \left(\frac{a - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}} \right) + \mu_{\tilde{y}} \\
&= \sigma_{\tilde{y}}\, \rho_{\tilde{a},\tilde{y}} \left(\frac{a - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}} \right) + \mu_{\tilde{y}}
\end{aligned}
$$

Uncorrelated features

Assuming the mean of the features and the response is zero

$$
\begin{aligned}
\ell_{\text{MMSE}}(\tilde{x}) &= \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} \tilde{x} \\
&= \begin{bmatrix} \sigma_{\tilde{x}[1],\tilde{y}} \\ \sigma_{\tilde{x}[2],\tilde{y}} \\ \cdots \\ \sigma_{\tilde{x}[d],\tilde{y}} \end{bmatrix}^T \begin{bmatrix} \sigma_{\tilde{x}[1]}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{\tilde{x}[2]}^2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \sigma_{\tilde{x}[d]}^2 \end{bmatrix}^{-1} \tilde{x} \\
&= \sum_{i=1}^{d} \frac{\sigma_{\tilde{x}[i],\tilde{y}}}{\sigma_{\tilde{x}[i]}^2} \tilde{x}[i] \\
&= \sum_{i=1}^{d} \frac{\sigma_{\tilde{y}} \rho_{\tilde{x}[i],\tilde{y}}}{\sigma_{\tilde{x}[i]}} \tilde{x}[i] \\
&= \sum_{i=1}^{d} \ell_{\text{MMSE}}(\tilde{x}[i])
\end{aligned}
$$

Noise cancellation

Goal: Estimate voice of a pilot $\tilde{y}$ from

$$\tilde{x}[1] = \tilde{y} + h\tilde{z} \quad \text{Inside helmet}$$

$$\tilde{x}[2] = h\tilde{y} + \tilde{z} \quad \text{Outside helmet}$$

$\tilde{z}$ is noise, independent from voice

Noise is much louder than voice: $\mathrm{Var}[\tilde{y}] = 1$, $\mathrm{Var}[\tilde{z}] = 100$

Everything is centered to have zero mean

Linear minimum MSE estimator

$$\ell_{\mathsf{MMSE}}(\tilde{x}) = \Sigma_{\tilde{x}\tilde{y}}^{T} \Sigma_{\tilde{x}}^{-1} \tilde{x}$$

Covariance matrix

$$\begin{aligned}
\mathrm{Var}[\tilde{x}[1]] &= \mathrm{Var}[\tilde{y} + h\tilde{z}] \\
&= \mathrm{Var}[\tilde{y}] + h^2 \mathrm{Var}[\tilde{z}] \\
&= 1 + 100h^2
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[\tilde{x}[2]] &= \mathrm{Var}[h\tilde{y} + \tilde{z}] \\
&= h^2 \mathrm{Var}[\tilde{y}] + \mathrm{Var}[\tilde{z}] \\
&= h^2 + 100
\end{aligned}$$

Covariance matrix

$$\begin{aligned}
\mathrm{Cov}[\tilde{x}[1], \tilde{x}[2]] &= \mathrm{E}\left[\tilde{x}[1]\tilde{x}[2]\right] \\
&= \mathrm{E}\left[\left(\tilde{y} + h\tilde{z}\right)\left(h\tilde{y} + \tilde{z}\right)\right] \\
&= \mathrm{E}\left[h\tilde{y}^2 + h\tilde{z}^2 + (1 + h^2)\tilde{y}\tilde{z}\right] \\
&= h\mathrm{E}\left[\tilde{y}^2\right] + h\mathrm{E}\left[\tilde{z}^2\right] + (1 + h^2)\mathrm{E}\left[\tilde{y}\right]\mathrm{E}\left[\tilde{z}\right] \\
&= 101h
\end{aligned}$$

Covariance matrix

$$\Sigma_{\tilde{x}} = \begin{bmatrix} 1 + 100h^2 & 101h \\ 101h & h^2 + 100 \end{bmatrix}$$

Cross-covariance

$$\begin{aligned}
\mathrm{Cov}[\tilde{x}[1], \tilde{y}] &= \mathrm{E}\left[\tilde{x}[1]\tilde{y}\right] \\
&= \mathrm{E}\left[\left(\tilde{y} + h\tilde{z}\right)\tilde{y}\right] \\
&= \mathrm{E}\left[\tilde{y}^2 + h\tilde{y}\tilde{z}\right] \\
&= \mathrm{E}\left[\tilde{y}^2\right] + h\mathrm{E}\left[\tilde{y}\right]\mathrm{E}\left[\tilde{z}\right] \\
&= 1 \\
\mathrm{Cov}[\tilde{x}[2], \tilde{y}] &= \mathrm{E}\left[\tilde{x}[2]\tilde{y}\right] \\
&= \mathrm{E}\left[\left(h\tilde{y} + \tilde{z}\right)\tilde{y}\right] \\
&= \mathrm{E}\left[h\tilde{y}^2 + \tilde{y}\tilde{z}\right] \\
&= h\mathrm{E}\left[\tilde{y}^2\right] + \mathrm{E}\left[\tilde{y}\right]\mathrm{E}\left[\tilde{z}\right] \\
&= h
\end{aligned}$$

$$\Sigma_{\tilde{x}\tilde{y}} = \begin{bmatrix} 1 \\ h \end{bmatrix}$$

Linear minimum MSE estimator

$$\begin{aligned}
\ell_{\text{MMSE}}(\tilde{x}) &= \Sigma_{\tilde{x}\tilde{y}}^{T} \Sigma_{\tilde{x}}^{-1} \tilde{x} \\
&= \begin{bmatrix} 1 & h \end{bmatrix} \begin{bmatrix} 1 + 100h^2 & 101h \\ 101h & h^2 + 100 \end{bmatrix}^{-1} \tilde{x} \\
&= \begin{bmatrix} 1 & h \end{bmatrix} \frac{1}{100(1-h^2)^2} \begin{bmatrix} h^2 + 100 & -101h \\ -101h & 1 + 100h^2 \end{bmatrix} \tilde{x} \\
&= \frac{1}{100(1-h^2)^2} \begin{bmatrix} 100(1-h^2) & -100h(1-h^2) \end{bmatrix} \tilde{x} \\
&= \frac{\tilde{x}[1] - h\tilde{x}[2]}{1-h^2}
\end{aligned}$$

How good is the estimator?

$$\begin{aligned}\ell_{\mathsf{MMSE}}(\tilde{x}) &= \frac{\tilde{x}[1] - h\tilde{x}[2]}{1-h^2} \\ &= \frac{\tilde{y} + h\tilde{z} - h(h\tilde{y} + \tilde{z})}{1-h^2} \\ &= \tilde{y}\end{aligned}$$

Gaussian random vectors

Gaussian random vector

$$\tilde{z} := \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \qquad \mu := \begin{bmatrix} \mu_{\tilde{x}} \\ \mu_{\tilde{y}} \end{bmatrix} \qquad \Sigma_{\tilde{z}} = \begin{bmatrix} \Sigma_{\tilde{x}} & \Sigma_{\tilde{x},\tilde{y}} \\ \Sigma_{\tilde{x},\tilde{y}}^T & \sigma_{\tilde{y}}^2 \end{bmatrix}$$

Conditional distribution of $\tilde{y}$ given $\tilde{x} = x$?

Gaussian with parameters

$$\mu_{\text{cond}} = \Sigma_{\tilde{x},\tilde{y}} \Sigma_{\tilde{x}}^{-1} (x - \mu_{\tilde{x}}) + \mu_{\tilde{y}}$$

$$\Sigma_{\text{cond}} = \Sigma_{\tilde{y}} - \Sigma_{\tilde{x},\tilde{y}} \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x},\tilde{y}}$$

Conditional mean of $\tilde{y}$ given $\tilde{x}$?

Gaussian random vectors

Gaussian random vector

$$\tilde{z} := \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \qquad \mu := \begin{bmatrix} \mu_{\tilde{x}} \\ \mu_{\tilde{y}} \end{bmatrix} \qquad \Sigma_{\tilde{z}} = \begin{bmatrix} \Sigma_{\tilde{x}} & \Sigma_{\tilde{x},\tilde{y}} \\ \Sigma_{\tilde{x},\tilde{y}}^T & \sigma_{\tilde{y}}^2 \end{bmatrix}$$

Conditional distribution of $\tilde{y}$ given $\tilde{x} = x$?

Gaussian with parameters

$$\mu_{\text{cond}} = \Sigma_{\tilde{x},\tilde{y}} \Sigma_{\tilde{x}}^{-1} (x - \mu_{\tilde{x}}) + \mu_{\tilde{y}}$$

$$\Sigma_{\text{cond}} = \Sigma_{\tilde{y}} - \Sigma_{\tilde{x},\tilde{y}} \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x},\tilde{y}}$$

Conditional mean of $\tilde{y}$ given $\tilde{x}$?

Linear estimation is optimal for Gaussians

Geometric intuition

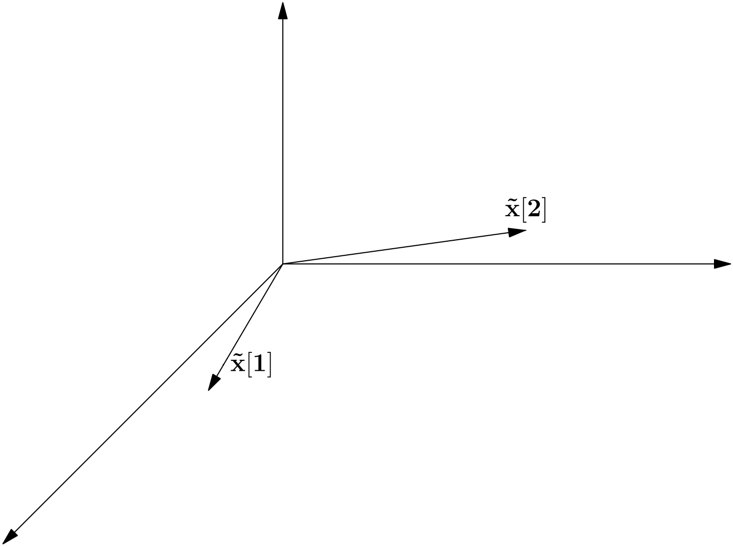
Zero-mean random variables can be interpreted as vectors

Inner product? Covariance

Squared norm / length? Variance

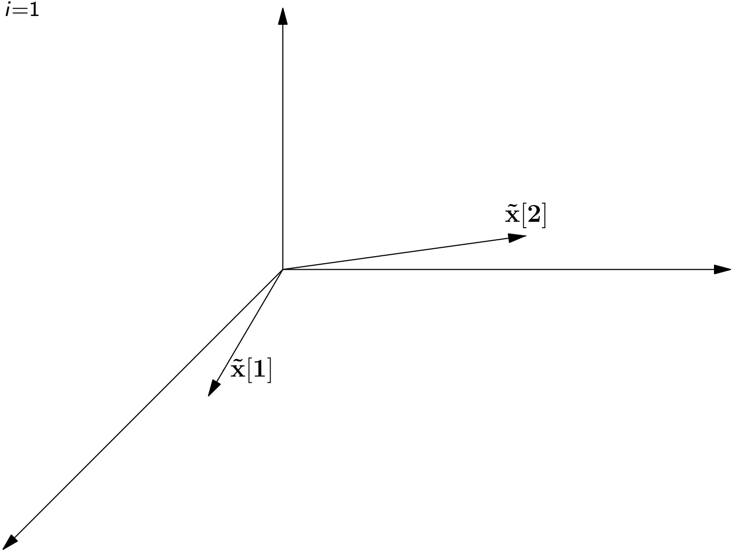
Mean-squared error? Squared distance

Features


$\tilde{\mathbf{x}}[\mathbf{2}]$
$\tilde{\mathbf{x}}[\mathbf{1}]$

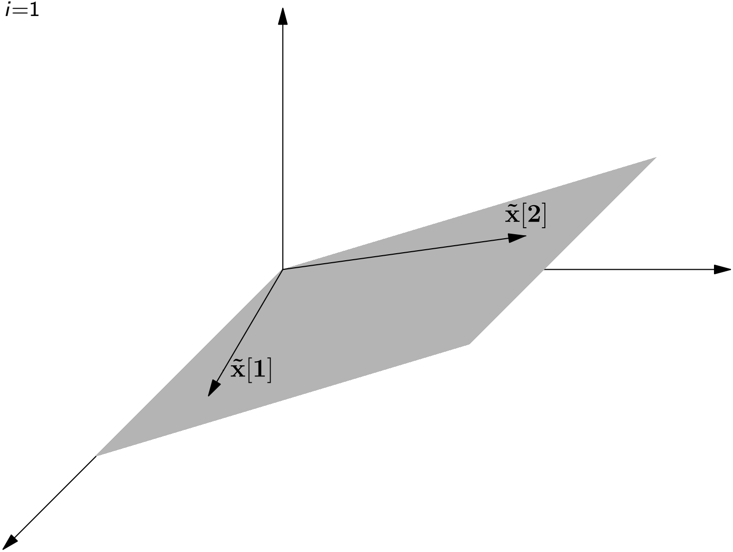
Linear combinations of features

$$\beta^T \tilde{x} = \sum_{i=1}^{d} \beta[i]\tilde{x}[i]$$


$\tilde{\mathbf{x}}[\mathbf{2}]$
$\tilde{\mathbf{x}}[\mathbf{1}]$

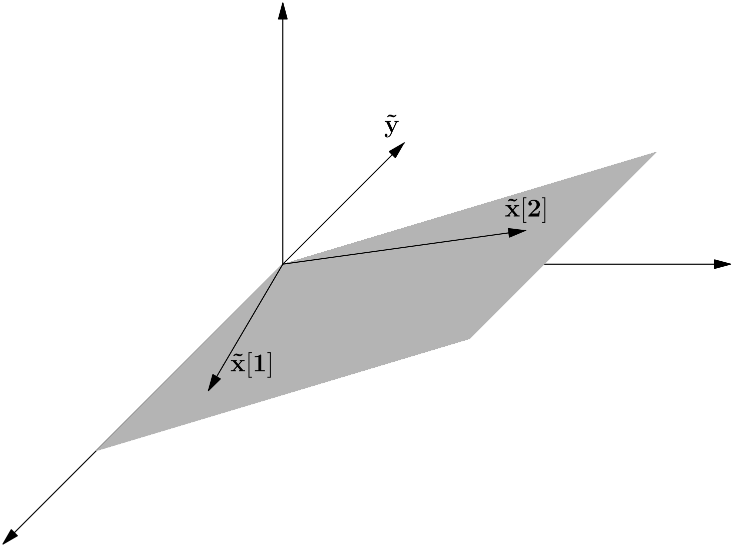
Linear combinations of features

$$\beta^T \tilde{x} = \sum_{i=1}^{d} \beta[i]\tilde{x}[i]$$


$\tilde{\mathbf{x}}[\mathbf{2}]$
$\tilde{\mathbf{x}}[\mathbf{1}]$

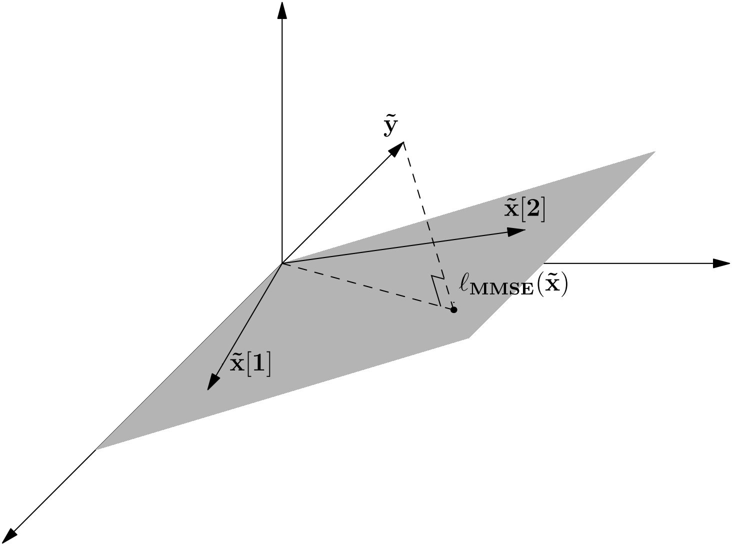
Response

$\beta^T \tilde{x}$ that is closest to $\tilde{y}$?


$\tilde{\mathbf{y}}$
$\tilde{\mathbf{x}}[\mathbf{2}]$
$\tilde{\mathbf{x}}[\mathbf{1}]$

Linear MMSE estimator

Orthogonal projection!

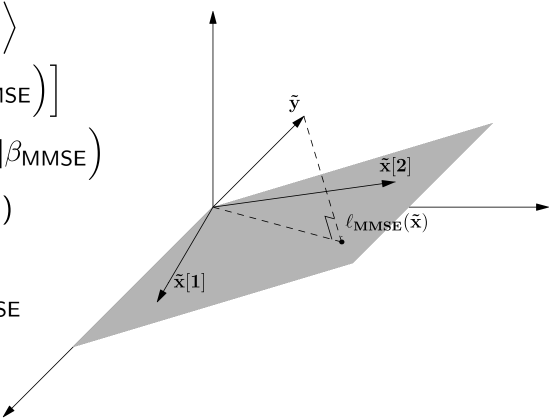

$\tilde{\mathbf{y}}$
$\tilde{\mathbf{x}}[\mathbf{2}]$
$\ell_{\mathbf{MMSE}}(\tilde{\mathbf{x}})$
$\tilde{\mathbf{x}}[\mathbf{1}]$

Orthogonal projection

$$\begin{aligned} 0 &= \left\langle \beta^T \tilde{x}, \tilde{y} - \ell_{\mathsf{MMSE}}(\tilde{x}) \right\rangle \\ &= \mathrm{E}\left[\beta^T \tilde{x}\left(\tilde{y} - \tilde{x}^T \beta_{\mathsf{MMSE}}\right)\right] \\ &= \beta^T \left(\mathrm{E}\left[\tilde{x}\tilde{y}\right] - \mathrm{E}[\tilde{x}\tilde{x}^T]\beta_{\mathsf{MMSE}}\right) \\ &= \beta^T \left(\Sigma_{\tilde{x}\tilde{y}} - \Sigma_{\tilde{x}}\beta_{\mathsf{MMSE}}\right) \end{aligned}$$

$$\Sigma_{\tilde{x}\tilde{y}} = \Sigma_{\tilde{x}}\beta_{\mathsf{MMSE}}$$

$$\beta_{\mathsf{MMSE}} = \Sigma_{\tilde{x}}^{-1}\Sigma_{\tilde{x}\tilde{y}}$$


$\tilde{\mathbf{y}}$
$\tilde{\mathbf{x}}[\mathbf{2}]$
$\ell_{\mathbf{MMSE}}(\tilde{\mathbf{x}})$
$\tilde{\mathbf{x}}[\mathbf{1}]$

What have we learned?

Derivation of linear minimum MSE estimator

Linear estimation is optimal for Gaussian random vectors

Geometric intuition