



Multivariate Linear Regression

Michael Reiter

Pattern Recognition and Image Processing Group, Vienna University of Technology,
Favoritenstr. 9, A-1040 Vienna, Austria
rei@prip.tuwien.ac.at, donner@prip.tuwien.ac.at, langs@prip.tuwien.ac.at

January 13, 2010



Consider two random vectors $\mathbf{x} \in \mathbf{R}^p$ and $\mathbf{y} \in \mathbf{R}^q$ with a joint probability

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x}). \quad (1)$$

Assume that \mathbf{x} and \mathbf{y} are related by

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) + \boldsymbol{\epsilon}, \quad (2)$$

where $\mathbf{f} : \mathbf{R}^p \rightarrow \mathbf{R}^q$ and $\boldsymbol{\epsilon} \in \mathbf{R}^q$ is a random noise vector with $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{Cov}(\boldsymbol{\epsilon}) = \Sigma$.



In this setting, the deterministic function $\mathbf{f}(\mathbf{x})$ is the mean of the conditional distribution of the output variables, conditioned on the input variables (see for example [Bishop 1995]), i.e.,

$$\mathbf{f}(\mathbf{x}) = E_{\mathbf{y}}(\mathbf{y}|\mathbf{x}) = \int \mathbf{y}p(\mathbf{y}|\mathbf{x})d\mathbf{y}. \quad (3)$$

Hence, if the joint probability density of \mathbf{x} and \mathbf{y} is known, we can determine the regression function \mathbf{f} by Eq. 3.

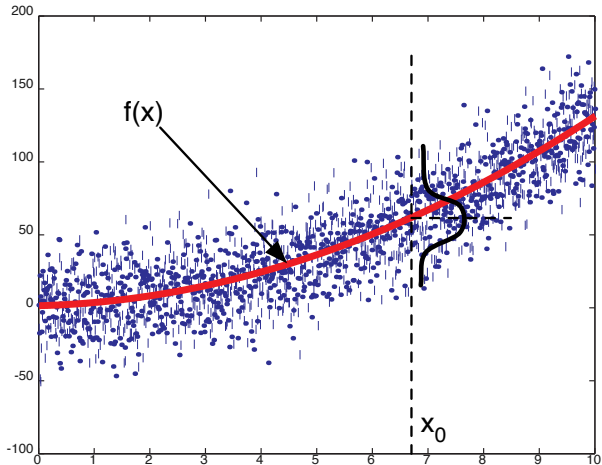


Figure: $f(\mathbf{x}_0) = E_y(\mathbf{y} | \mathbf{x} = \mathbf{x}_0)$.



- We are only given a sample $\mathcal{T} = \{\mathbf{x}_i, \mathbf{y}_i, i = 1, \dots, N\}$ of pairs of corresponding observations of the random variables \mathbf{x} and \mathbf{y} .
- A model (e.g., a neural network) capable of implementing a set of functions \mathcal{F} is used.
- Goal: find the optimal approximating function $\hat{\mathbf{f}}^* = \hat{\mathbf{f}}_{w^*} \in \mathcal{F}$



- How assess the quality of the approximating function $\hat{\mathbf{f}}_w$?
- Use a *loss function* $L(\mathbf{y}, \hat{\mathbf{f}}_w(\mathbf{x}))$ (defined pointwise)

Example (Squared Loss)

$$L(\mathbf{y}, \hat{\mathbf{f}}_w(\mathbf{x})) = \|\mathbf{y} - \hat{\mathbf{f}}_w(\mathbf{x})\|^2 \quad (4)$$



Log-likelihood loss

General loss: negative log-likelihood of the response density of \mathbf{y} at a given \mathbf{x} , i.e.

$$L(\mathbf{y}, \theta(\mathbf{x})) = -2 \log p_{\theta(\mathbf{x})}(\mathbf{y}), \quad (5)$$

where θ is a parameter of a probability density depending (conditioned) on \mathbf{x} .

Example (additive Gaussian error)

$$p_{\theta(\mathbf{x})}(\mathbf{y}) = N(\hat{\mathbf{f}}_w(\mathbf{x}), \Sigma). \quad (6)$$



Minimizing risk

- The expected value of the loss is called the (*overall*) risk

$$R(\hat{\mathbf{f}}_w) = E_{\mathbf{x}}E_{\mathbf{y}|\mathbf{x}}L(\mathbf{y}, \hat{\mathbf{f}}_w(\mathbf{x})). \quad (7)$$

- The optimal approximating function is given by

$$\hat{\mathbf{f}}_{w^*} = \arg \min_{\hat{\mathbf{f}}_w \in \mathcal{F}} R(\hat{\mathbf{f}}_w). \quad (8)$$

- If $\mathbf{f} \in \mathcal{F}$, then is easy to show that $\hat{\mathbf{f}}_{w^*} = \mathbf{f}$



Empirical Risk Minimization

- We seek an estimate of the function \mathbf{f} based on a sample \mathcal{T} of N of observations (realizations).
- Typically, the *training error* rate (sometimes referred to as *empirical risk*) is used:

$$R_{emp}(\hat{\mathbf{f}}_w) = \frac{1}{N} \sum_{i=1}^N L(\mathbf{y}_i, \hat{\mathbf{f}}_w(\mathbf{x}_i)) \quad (9)$$



Least Squares Estimator

- The most common estimation method is minimizing the training error with squared error loss which leads to the residual sum-of-squares error function (RSS)

$$\text{RSS}(\hat{\mathbf{f}}_w) = \frac{1}{N} \sum_{i=1}^N (\mathbf{y}_i - \hat{\mathbf{f}}_w(\mathbf{x}_i))^2.$$

- This leads to the *least-squares estimator*

$$\hat{\mathbf{f}}_{w^*} = \arg \min_{\hat{\mathbf{f}}_w \in \mathcal{F}} \text{RSS}(\hat{\mathbf{f}}_w).$$



The estimator $\hat{\mathbf{f}}(\mathbf{x}) = \hat{\mathbf{f}}_{w^*}(\mathbf{x})$ of the output at an arbitrary position \mathbf{x} is random as a function of the observed sample \mathcal{F} .



Expected prediction error (EPE)

The *expected prediction error* (EPE) corresponds to the *expected* loss of the prediction $\hat{\mathbf{f}}(\mathbf{x})$, i.e.,

$$\text{EPE}(\hat{\mathbf{f}}(\mathbf{x})) = E_{\mathcal{T}} E_{\mathbf{y}|\mathbf{x}} L(\mathbf{y}, \hat{\mathbf{f}}(\mathbf{x})) \quad (10)$$

where $E_{\mathcal{T}}$ denotes the expectation over all possible samples.



EPE vs. MSE

If we use squared error loss we can decompose the EPE as

$$\begin{aligned}
 \text{EPE}(\hat{\mathbf{f}}(\mathbf{x})) &= E_{\mathcal{T}} E_{\mathbf{y}|\mathbf{x}}(\|\mathbf{y} - \hat{\mathbf{f}}(\mathbf{x})\|^2) \\
 &= \underbrace{E_{\mathcal{T}}(\|\mathbf{f}(\mathbf{x}) - \hat{\mathbf{f}}(\mathbf{x})\|^2)}_{\text{mean squared error (MSE)}} + \underbrace{\text{trace}(\Sigma)}_{\text{irreducible}}. \quad (11)
 \end{aligned}$$



Overall Error

$$\text{EPE}_{\hat{\mathbf{f}}} = \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathcal{T}} \mathbb{E}_{\mathbf{y}|\mathbf{x}} L(\mathbf{y}, \hat{\mathbf{f}}(\mathbf{x})).$$



Model Complexity, Bias, Variance

The MSE is of particular importance because it can be recast as

$$\text{MSE}(\hat{\mathbf{f}}(\mathbf{x})) = \underbrace{\|\mathbb{E}_{\mathcal{T}}\hat{\mathbf{f}}(\mathbf{x}) - \mathbf{f}(\mathbf{x})\|^2}_{\text{Bias}^2(\hat{\mathbf{f}}(\mathbf{x}))} + \underbrace{\mathbb{E}_{\mathcal{T}}\|\hat{\mathbf{f}}(\mathbf{x}) - \mathbb{E}_{\mathcal{T}}\hat{\mathbf{f}}(\mathbf{x})\|^2}_{\text{Var}(\hat{\mathbf{f}}(\mathbf{x}))}. \quad (12)$$



Linear Regression

The linear regression model assumes that \mathbf{f} has the form (or can be approximated by)

$$\mathbf{f}(\mathbf{x}) = E(\mathbf{y}|\mathbf{x}) = \mathbf{W}\mathbf{x} + \mathbf{w}_0, \quad (13)$$



Gaussian setting

We assume a stationary ergodic environment in which \mathbf{x} and \mathbf{y} are jointly gaussian, such that the environment can be described by the second-order statistics

- $\mathbf{C}_{xx} = E(\mathbf{x}\mathbf{x}^T)$, which is the covariance of \mathbf{x} and
- $\mathbf{C}_{xy} = E(\mathbf{x}\mathbf{y}^T)$, the cross-covariance of \mathbf{x} and \mathbf{y} and $\mathbf{C}_{yx} = \mathbf{C}_{xy}^T$,



Wiener filter

The coefficients \mathbf{W} are given by the *Wiener filter solution*

$$\mathbf{W} = \mathbf{C}_{yx} \mathbf{C}_{xx}^{-1}. \quad (14)$$

to the linear optimum filtering problem.



Wiener filter

The Wiener solution corresponds to the *least mean square solution* in the sense that the expected error (risk) reaches its minimum:

$$R(\mathbf{f}_w) = E_x E_{y|x} L(\mathbf{y}, \mathbf{f}_w(\mathbf{x})) \quad (15)$$

$$= \text{trace}(\mathbf{C}_{yy} - E(\mathbf{f}_w(\mathbf{x})\mathbf{f}_w(\mathbf{x})^T)) \quad (16)$$

$$= \text{trace}(\mathbf{C}_{yy} - \mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}) \quad (17)$$

$$= \text{trace}(E(\epsilon\epsilon^T)) = q\sigma^2, \quad (18)$$

where \mathbf{f}_w denotes the true linear model



Estimates

We seek an estimate of the true parameters \mathbf{W} minimizing the residual sum-of-squares error criterion, i.e.,

$$\hat{\mathbf{W}} = \arg \min \text{RSS}(\mathbf{W})$$

where

$$\begin{aligned} \text{RSS}(\mathbf{W}) &= \sum_{i=1}^N (\mathbf{y}_i - \mathbf{f}_{\mathbf{w}}(\mathbf{x}_i))^2 \\ &= \sum_{i=1}^N (\mathbf{y}_i - \mathbf{W}\mathbf{x}_i)^T (\mathbf{y}_i - \mathbf{W}\mathbf{x}_i) \\ &= \text{trace} \left((\mathbf{Y} - \mathbf{W}\mathbf{X})^T (\mathbf{Y} - \mathbf{W}\mathbf{X}) \right). \end{aligned} \quad (19)$$



Estimates

The estimator $\hat{\mathbf{W}}$ is obtained by setting the derivative of Eq. 19 to zero and is given by

$$\hat{\mathbf{W}} = \mathbf{YX}^T(\mathbf{XX}^T)^{-1}. \quad (20)$$



Linear Predictions

The predicted values for the training data are

$$\hat{\mathbf{Y}} = \hat{\mathbf{W}}\mathbf{X} = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}, \quad (21)$$

where the i -th column of $\hat{\mathbf{Y}}$ is $\hat{\mathbf{y}}_i = \hat{\mathbf{W}}\mathbf{x}_i$. The matrix $\mathbf{H} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}$ in the above equation is called the "hat" matrix because it puts a hat on \mathbf{Y} . The matrix \mathbf{H} corresponds to a projection onto the row space of \mathbf{X} .



Approximations of the prediction error

- It is impossible to determine the EPE of an estimator without knowledge of the densities of \mathbf{x} and \mathbf{y} .
- However, we can obtain an approximation as follows: Let

$$\mathbf{F} = E_{\mathbf{x}}(\mathbf{f}(\mathbf{x})\mathbf{f}(\mathbf{x})^T) \quad (22)$$

$$= \mathbf{C}_{xy}^T \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \quad (23)$$

$$= \mathbf{C}_{yy} - \Sigma \quad (24)$$



let $\mathbf{h}(\mathbf{x}) = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{x}$ so that $\hat{\mathbf{f}}(\mathbf{x}) = \mathbf{Y}\mathbf{h}(\mathbf{x})$. Then, if we condition on the design \mathbf{X} and assume that only ϵ is random, we can write

$$\begin{aligned} E_{\mathbf{Y}|\mathbf{X}} \left[\frac{1}{N} \sum_{i=1}^N \hat{\mathbf{f}}(\mathbf{x}_i) \hat{\mathbf{f}}(\mathbf{x}_i)^T \right] &= \frac{1}{N} E_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}\mathbf{H}\mathbf{Y}^T) \\ &= \frac{p}{N} \mathbf{\Sigma}. \end{aligned} \quad (25)$$



Assuming that the sample mean and covariance of the input observations \mathbf{x}_i are equal to the true mean and covariance, as a consequence of Eq. 25 we can write

$$E(\hat{\mathbf{f}}(\mathbf{x})\hat{\mathbf{f}}(\mathbf{x})^T) = \frac{p}{N}\mathbf{\Sigma} + \mathbf{F} \quad (26)$$

$$= \mathbf{C}_{yy} + \left(\frac{p}{N} - 1\right)\mathbf{\Sigma} \quad (27)$$