Regression

Anita Faul

Laboratory for Scientific Computing, University of Cambridge

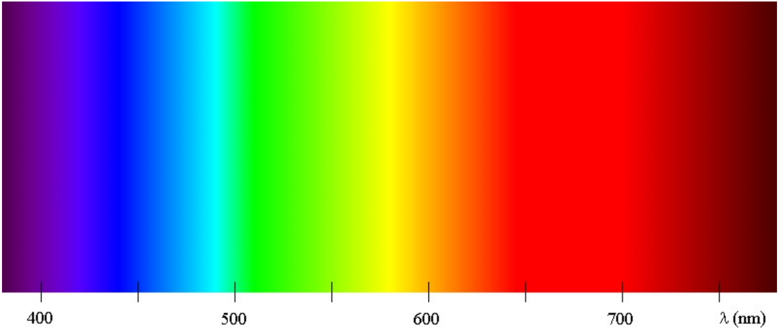
Regression

Regression: Estimating the relationship between variables. It is related to:

- Curve fitting,
- Interpolation,
- Data Prediction.

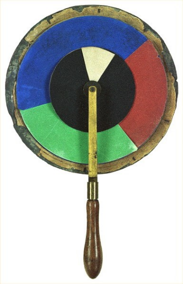
Vision

Spectral colours are evoked by a single wavelength. Colours change continuously.

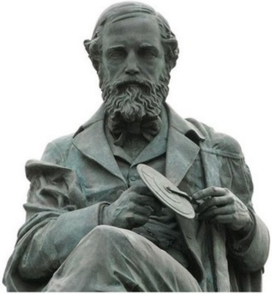


A physical colour is a combination of pure spectral colours. Infinitely many possibilities.

Maxwell's Colour Top

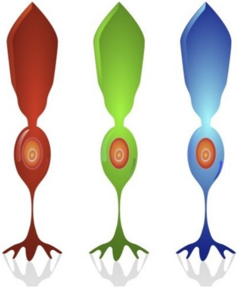


- Three coloured paper discs overlapping by different amounts.
- Smaller central disc with wedge of colour sample.
- All colours can be mixed by different combinations of the three primary colours, red, green and blue.



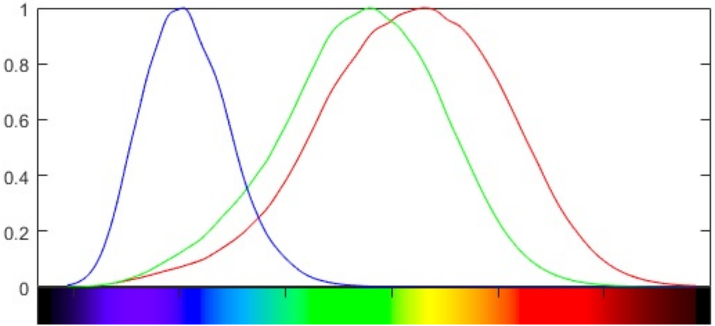
Vision

The human eye has only three measuring devices to perceive colours – the cones.



Vision

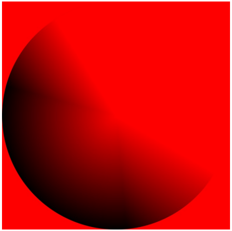
Sensitivity of cones to different wavelengths:



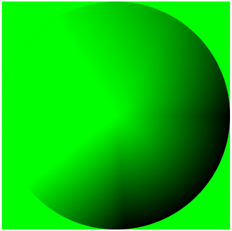
With three measurements, the human eye can distinguish 10 million colours.

Vision

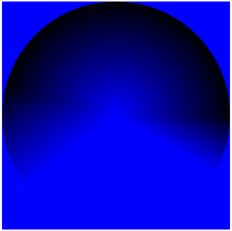
Not just spectral colours but also non-spectral colours.



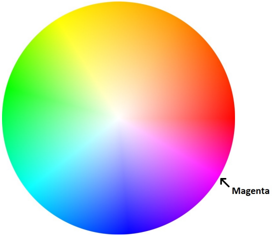
+



+



=



Regression in action.

Magenta

Magenta

- Invented in 1859/60 by François-Emmanuel Verguin (Lyon), Chambers Nicolson and George Maule (London),
- anilin dye,
- named after the battle of Magenta,
- mixture of the primary colours red and blue,
- complementary color to green,
- thus also generated when the green component of light is absorbed.

Dress by Madame Vignon, c1870

Multichromacy

Chromacy	Types of cone cells	Approximate number of colours perceived	Examples
Monochromacy	1	100	Marine mammals
Dichromacy	2	10,000	Most terrestrial mammals
Trichromacy	3	10 million	Humans, great apes, some insects
Tetrachromacy	4	100 million	Most reptiles, amphibians, birds and insects, rarely humans
Pentachromacy	5	10 billion	Some insects (butterflies), some birds (pigeons).

We cannot imagine how a butterfly sees the world.

Linearisation

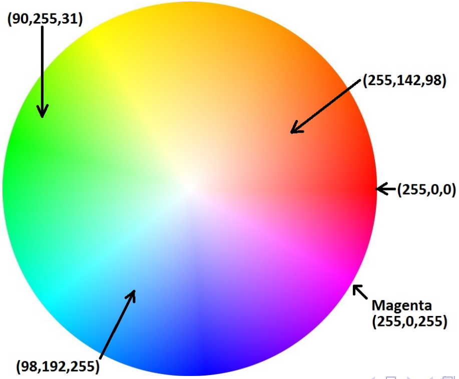
- There are three base colours: red, green, blue.
- All other colours are mixed from these.
- They are represented in an array of bytes taking values between 0 and 255: $[r, g, b]$.
- red $= (255, 0, 0)$.
- green $= (0, 255, 0)$.
- blue $= (0, 0, 255)$.

Linearisation

- Linearisation means multiplications by a number and additions are allowed.
- red $+$ green $= (255, 255, 0) =$ yellow.
- red $+ 0.5*$ green $= (255, 128, 0) =$ orange.
- red $+ 0.5*$ blue $= (255, 0, 128) =$ pink.
- $0.6*$ red $+ 0.8*$ blue $= (153, 0, 204) =$ purple.

Regression

- Regression: Estimating the relationship between variables.
- Given the intensity of red, green, blue for some samples, what is the relationship between intensity values and the colour?



Applications

Applications

- Computer Vision: e.g. image compression and restoration,
- Engineering: e.g. machine degradation,
- Medicine: e.g. epidemiology, mammography,
- Finance: e.g. volatility prediction, pricing models,
- Econometrics: e.g. cost and benefit optimization,
- Hydrology: e.g. flow prediction, ground water level forecasting,
- Seismology: e.g. soil liquefaction, seismic surveys.
- Suggestions?

Problem description

- Given measurements $t_1, \ldots, t_N$, each measurement depends on parameters we know $\mathbf{x}_1, \ldots, \mathbf{x}_N$. The intensities of red, green, blue.
- These are quantities which can be measured with more or less effort.
- For example by photocells, geophones, hydrophones, PET and MRI neuroimaging, EEG technology, etc.
- In the following it is assumed that the measurements are mean centred $(1/N(t_1 + \cdots + t_N) = 0)$. This eliminates the need for a bias in the model.

Unknowns

- The measurements also depend on parameters we do not know.
- Any real world application depends on factors which cannot be measured.
- Or these measurements would be disproportionally difficult, costly or invasive.

Processes

- We assume that the measurements are the result of underlying processes following some laws.
- In some applications, these laws are known, and for example described by partial differential equations but the specific parameters are not known.
- Sometimes nothing is known and we have to try to find this out.

Dictionaries

- If we had a solution to the underlying process, we could predict the measurement from a function $t(\mathbf{x})$ as

  $$t_n = t(\mathbf{x}_n),$$

  where parameters of the function depend on the process.
- If we had a set of candidate functions $d_1(\mathbf{x}), \ldots, d_M(\mathbf{x})$, we could try which fits the measurements and thus infer the underlying process.
- We say the functions $d_1(\mathbf{x}), \ldots, d_M(\mathbf{x})$ form a dictionary and let

  $$f(\mathbf{x}) = \sum_{m=1}^{M} c_m d_m(\mathbf{x}),$$

  be an approximation to $t(\mathbf{x})$, where $c_1, \ldots, c_M$ are coefficients and these need to be determined.

Linear Regression

- The term linear means that the model is a sum of building blocks with suitable coefficients.
- Flexibility in the choice of building blocks. They can be non-linear functions.
- Flexibility in the choice of noise model.
- Flexibility in the method to determine the coefficients.

Linear Regression - Terminology

Linear regression - Terminology

- Scalar function value t, also known as *dependent, endogenous, response, measured, criterion variable* or *regressand*. The colour in our example.
- $\mathbf{d} = (d_1(\mathbf{x}), \ldots, d_M(\mathbf{x}))^T$ *independent, exogenous, input, explanatory, predictor variables* or *regressors*.
- Remember d_j is NOT a coordinate of a vector. It is a basis function.
- $\mathbf{c} = (c_1, \ldots, c_M)^T$ *parameter vector* or *vector of weights*, also known as *effects, regression coefficients*. They can also be viewed as *latent* variables. Regression determines these by various techniques.

Noise

- The relationship to the measurements is

$$t_n = f(\mathbf{x}_n) + \epsilon_n,$$

- where ϵ_n is noise intrinsic to the measurement process
- and assumed to be independent and identically, normally distributed, $\mathcal{N}(0, \sigma^2)$ (*homoscedasticity*).
- Note: The assumption of the same constant variance might be wrong:
    - There might be different sources of the noise with different effects. In this case the noise should be modeled by a mixture of probability distributions.
    - Restrictions on the data might make different noise distributions necessary. For example if it is known that the data is always positive, the noise variance for smaller values has to be smaller to ensure positive predicted values.

Mathematical model

- Now

$$t_n = f(\mathbf{x}_n) + \epsilon_n = \sum_{m=1}^{M} c_m d_m(\mathbf{x}_n) + \epsilon_n,$$

- Let $\mathbf{D}$ be the matrix with entries

$$D_{n,m} = d_m(\mathbf{x}_n)$$

and let $\mathbf{t}^T = (t_1, \ldots, t_N)$, $\mathbf{c}^T = (c_1, \ldots, c_M)$ and $\boldsymbol{\epsilon}^T = (\epsilon_1, \ldots, \epsilon_N)$, then

$$\mathbf{t} = \mathbf{D}\mathbf{c} + \boldsymbol{\epsilon}.$$

Predictions

- The matrix $\mathbf{D}$ is called the *design matrix*.
- The challenge is to find the dictionary of basis functions and the coefficients.
- Once these are found, predictions for new data $\mathbf{x}$ can be made by

$$t = \sum_{m=1}^{M} c_m d_m(\mathbf{x}).$$

- Or, by defining $\mathbf{d}(\mathbf{x})^T = (d_1(\mathbf{x}), \dots, d_M(\mathbf{x}))$,

$$t = \mathbf{d}(\mathbf{x})^T \mathbf{c}.$$

Polynomial Regression

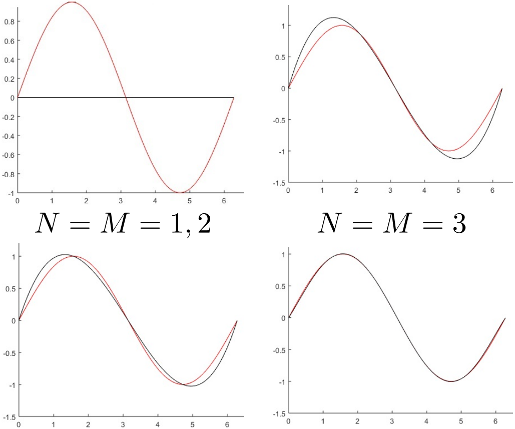
- Dating back to Lagrange (1805) and Gauss (1809).
- y is modeled as a polynomial of degree $M-1$.
- $1, x, \ldots, x^{M-2}, x^{M-1}$ form a basis of polynomials of degree $M-1$, our dictionary.
- The design matrix is

$$\mathbf{D} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^{M-2} & x_1^{M-1} \\ \vdots & \vdots & \ddots & \ldots & \vdots \\ 1 & x_N & \cdots & x_N^{M-2} & x_N^{M-1} \end{pmatrix}$$

- Assuming no noise, the solution is unique if $N = M$.

Polynomial Regression

Approximating $y = sin(x)$.



$N = M = 1, 2$

$N = M = 3$

$N = M = 4$

$N = M = 5$

Polynomial Regression

- Could we have been more clever in the design phase?
- $\sin(0) = 0$.
- $\sin(x)$ is an odd function, that is $\sin(-x) = -\sin(x)$.
- Moreover, $\sin(x) = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + \ldots$.
- A better choice might have been

$$\mathbf{D} = \begin{pmatrix} x_1 & x_1^3 & \cdots & x_1^{2M-3} & x_1^{2M-1} \\ \vdots & \vdots & \ddots & \ldots & \vdots \\ x_N & x_N^3 & \cdots & x_N^{2M-3} & x_N^{2M-1} \end{pmatrix}$$

- Still M regressors, but possibly higher accuracy.
- *The choice of regressors depends on the problem.*

Ordinary Least Squares (OLS)

- Assume the dictionary of basis functions is known.
- How best to solve $\mathbf{t} = \mathbf{D}\mathbf{c} + \boldsymbol{\epsilon}$ generally?
- Remember the ϵ_i are independent and identical normally distributed random variables with zero mean and variance σ^2.
- The *likelihood* of observing $\mathbf{t}$ given the model specified by $\mathbf{D}$, $\mathbf{c}$ and σ^2 is

$$\mathcal{L}(\mathbf{t}|\mathbf{D},\mathbf{c},\sigma^2) = (2\pi\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{t}-\mathbf{D}\mathbf{c})^T(\mathbf{t}-\mathbf{D}\mathbf{c})\right).$$

- The *log likelihood* is

$$\log\mathcal{L}(\mathbf{t}|\mathbf{D},\mathbf{c},\sigma^2) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{t}-\mathbf{D}\mathbf{c})^T(\mathbf{t}-\mathbf{D}\mathbf{c}).$$

OLS

- $r_n = t_n - (d_1(\mathbf{x}_n), \dots, d_M(\mathbf{x}_n))^T \mathbf{c} = t_n - \mathbf{d}(\mathbf{x}_n)^T \mathbf{c}$ is called the *n-th residual*.
- Note that $\mathbf{d}(\mathbf{x}_n)^T$ is the n-the row of $\mathbf{D}$.
- The *sum of squared residual (SSR)* (also known as *error sum of squares (ESS)* or *residual sum of squares (RSS)*) is

$$\sum_{n=1}^{N} r_n^2 = \sum_{n=1}^{N} (t_n - \mathbf{d}(\mathbf{x}_n)^T \mathbf{c})^2 = (\mathbf{t} - \mathbf{Dc})^T (\mathbf{t} - \mathbf{Dc}) = \|\mathbf{t} - \mathbf{Dc}\|^2.$$

- Maximizing the log likelihood is equivalent to minimizing this sum.
- This sum is minimal if

$$\mathbf{c} = (\mathbf{D}^T \mathbf{D})^{-1} \mathbf{D}^T \mathbf{t}.$$

OLS

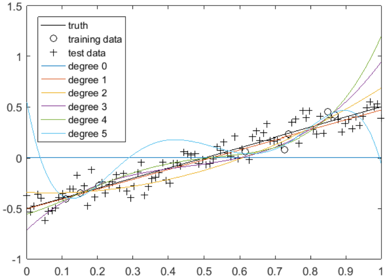
- For $\mathbf{x}_1, \ldots, \mathbf{x}_N$, the predicted values $\hat{\mathbf{t}} = (\hat{t}_1, \ldots, \hat{t}_N)^T$ are given by

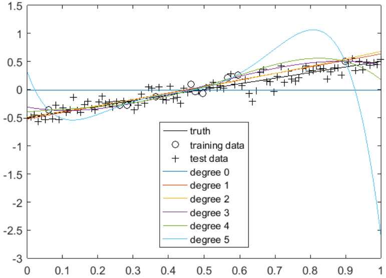
$$\hat{\mathbf{t}} = \mathbf{D}\mathbf{c} = \mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{t}.$$

- Geometrically, this is the projection of $\mathbf{t}$ onto the space spanned by the columns of $\mathbf{D}$.
- If $\mathbf{t}$ already lies in that space, then $\hat{\mathbf{t}} = \mathbf{t}$. In this case, the regressors are perfectly suited to predict the data.
- For an unseen $\mathbf{x}$, t can then be predicted by $t = \mathbf{c}^T\mathbf{d}(\mathbf{x})$.

Under-fitting and Over-fitting

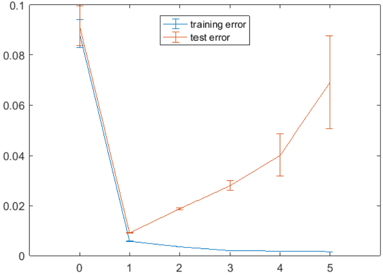
- Under-fitting occurs, if there are not enough explanatory variables to explain the data.
- Over-fitting occurs, if the model also models the noise.
- Another problem are unsuitable regressors.

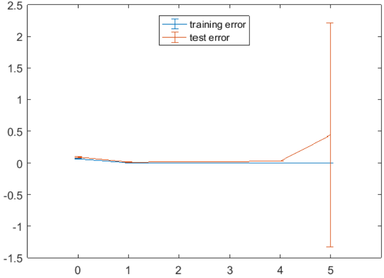




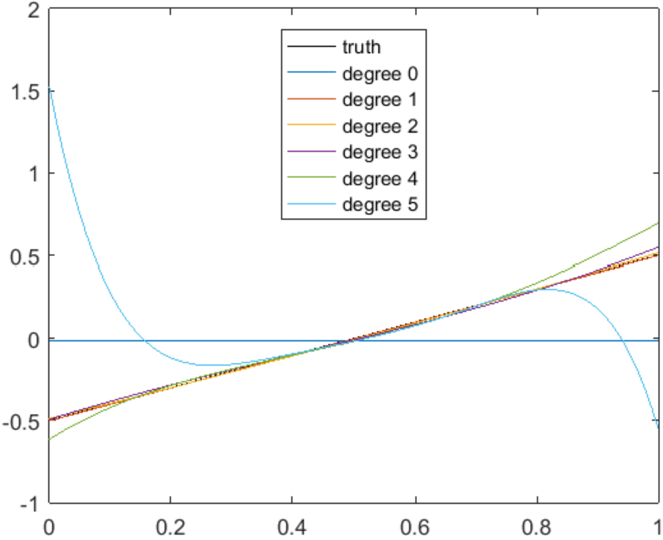
Under-fitting and Over-fitting

Whether a model under-fits or over-fits, can be examined by considering the mean squared test and training error.





Average Polynomial



Average Polynomial

Coefficients of the average polynomials:

degree	x^5	x^4	x^3	x^2	x	1
0	0	0	0	0	0	-0.011
1	0	0	0	0	1.0	-0.501
2	0	0	0	0.079	0.92	-0.49
3	0	0	0.50	-0.61	1.1	-0.49
4	0	-1.5	4.8	-4.2	2.3	-0.61
5	-63	163	-164	81	-19	1.5

Average Polynomial

Variance of the coefficients:

degree	x^5	x^4	x^3	x^2	x	1
0	0	0	0	0	0	0.0092
1	0	0	0	0	0.011	0.0035
2	0	0	0	0.42	0.51	0.033
3	0	0	0.50	31.9	9.03	0.3
4	0	1227	5030	2869	324.1	5.36
5	338000	2016000	1829000	395848	20873	182.1

Bias and Variance

Given a model space, let $f(\mathbf{x})$ be the approximation to the underlying process $t(\mathbf{x})$ generating the data. Note $f(\mathbf{x})$ depends on the data set used to calculate the model. The measurement is $t = t(\mathbf{x}) + \epsilon$. The expected squared error at $\mathbf{x}$ is:

$$\begin{aligned}
\mathrm{E}\left[(t - f(\mathbf{x}))^2\right] &= \mathrm{E}\left[t^2 + f(\mathbf{x})^2 - 2tf(\mathbf{x})\right] \\
&= \mathrm{E}\left[t^2\right] + \mathrm{E}\left[f(\mathbf{x})^2\right] - 2\mathrm{E}\left[tf(\mathbf{x})\right] \\
&= \mathrm{Var}\left[t\right] + \mathrm{E}\left[t\right]^2 + \mathrm{Var}\left[f(\mathbf{x})\right] + \mathrm{E}\left[f(\mathbf{x})\right]^2 - 2\mathrm{E}\left[tf(\mathbf{x})\right] \\
&= \sigma^2 + t(\mathbf{x})^2 + \mathrm{Var}\left[f(\mathbf{x})\right] + \mathrm{E}\left[f(\mathbf{x})\right]^2 - 2t(\mathbf{x})\mathrm{E}\left[f(\mathbf{x})\right] \\
&= \sigma^2 + \underbrace{\mathrm{Var}\left[f(\mathbf{x})\right]}_{\textit{variance}} + \underbrace{(\mathrm{E}\left[f(\mathbf{x})\right] - t(\mathbf{x}))^2}_{\textit{bias}}.
\end{aligned}$$

Bias and Variance

	low variance	high variance
low bias	model space adequate and robust to changes in training data	model susceptible to changes in training data
high bias	model space inadequate for underlying process	undesirable

In general, making one smaller, increases the other, known as *bias-variance trade-off*.

Cross-validation

- *Leave-p-out cross-validation* sets aside p samples for validation. All possibilities to choose p validation samples from n samples are considered. Thus the regression algorithm is run $\frac{n!}{p!(n-p)!}$ times.
- *Leave-one-out cross-validation* where $p = 1$. The algorithm is run n times.
- *k-fold cross-validation* randomly subdivides the data set into k equal sized subsets. One of these is the validation set while the other $k - 1$ form the training set. The model generation is repeated k times (the *folds*) with each of the subsets being the validation set exactly once. The k results are averaged for assessment. When $k = n$, k-fold cross-validation is the same as leave-one-out cross-validation.

Multicollinearity

- *Multicollinearity* happens when one or more predictor variables are highly correlated.
- In this case one of the predictors can be modeled by the others.
- A high degree of correlation increases the variance in the coefficients, since different models are equivalent.
- Small changes in the input data can lead to large changes in the model.
- *Perfect multicollinearity* means the regressors are linearly dependent, that is one can be exactly expressed by the others.
- In this case $(\mathbf{D}^T\mathbf{D})^{-1}$ does not exist.

Multicollinearity

- In a good regression model, the regressors correlate minimally with each other, but are each highly correlated with the regressand.
- The aim is to find a linear combination of few regressors which summarize and explain the data without too much loss of information.
- *Principal component regression* uses the principal components of $\mathbf{D}$ as regressors instead of the columns of $\mathbf{D}$.

Principal Component Regression

- Let $\mathbf{d}_1, \ldots, \mathbf{d}_M$ denote the columns of $\mathbf{D}$.
- We assume that the columns are standardized that is they have mean 0 and length 1.
- The first assumption is valid, since the measurements are mean centred, and the second assumption is valid, since regressors are invariant to scaling.
- The correlation between the i-th and j-th regressor is then

$$\mathbf{d}_i^T \mathbf{d}_j.$$

- The correlation matrix is $\mathbf{D}^T\mathbf{D}$.

Principal Component Regression

- A new set of regressors is generated as linear combinations of regressors, say

$$v_1 d_1(\mathbf{x}) + \ldots + v_M d_M(\mathbf{x}).$$

- Evaluating this at the N different points we arrive at a linear combination of the columns of $\mathbf{D}$,

$$v_1 \mathbf{d}_1 + \ldots + v_M \mathbf{d}_M = \mathbf{D}\mathbf{v}.$$

- The correlation between two such linear combinations is

$$\frac{\mathbf{v}_1^T \mathbf{D}^T \mathbf{D} \mathbf{v}_2}{\|\mathbf{D}\mathbf{v}_1\| \|\mathbf{D}\mathbf{v}_2\|},$$

where $\mathbf{v}_1$ and $\mathbf{v}_2$ are such that $\mathbf{D}\mathbf{v}_1 \neq 0$ and $\mathbf{D}\mathbf{v}_2 \neq 0$.

Principal Component Regression

- The matrix $\mathbf{D}^T\mathbf{D}$ is symmetric and positive semidefinite.
- It has M non-negative eigenvalues and corresponding orthonormal eigenvectors.
- Thus choosing $\mathbf{v}_1$ and $\mathbf{v}_2$ to be eigenvectors, the new regressors are uncorrelated.
- The eigenvectors corresponding to the K nonzero eigenvalues are chosen. The prediction $\hat{\mathbf{t}}$ is the projection of $\mathbf{t}$ onto the space spanned by $\mathbf{D}\mathbf{v}_1, \ldots, \mathbf{D}\mathbf{v}_K$.
- The distance is

$$\|\mathbf{t} - \hat{\mathbf{t}}\|^2 = \mathbf{t}^T\mathbf{t} - \sum_{i=1}^{K} \frac{\left[(\mathbf{D}\mathbf{v}_i)^T\mathbf{t}\right]^2}{\lambda_k}.$$

Principal Component Regression

- We decompose the regressand $\mathbf{t}$ into one portion lying in the subspace spanned by $\mathbf{D}\mathbf{v}_1, \ldots, \mathbf{D}\mathbf{v}_K$ and a remainder $\mathbf{a}$ which is orthogonal to $\mathbf{D}\mathbf{v}_1, \ldots, \mathbf{D}\mathbf{v}_K$, $\mathbf{t} = \sum_{k=1}^{K} a_k \mathbf{D}\mathbf{v}_k + \mathbf{a}$.
- Then $(\mathbf{D}\mathbf{v}_i)^T \mathbf{t} = \sum_{k=1}^{K} a_k \mathbf{v}_i^T \mathbf{D}^T \mathbf{D}\mathbf{v}_k + \mathbf{v}_i^T \mathbf{D}^T \mathbf{a} = a_i \lambda_i$,
- $\|\mathbf{t} - \hat{\mathbf{t}}\|_2^2 = \mathbf{t}^T \mathbf{t} - \sum_{i=1}^{K} a_i^2 \lambda_i = \|\mathbf{a}\|_2^2$.
- If sparsity is required and not all eigenvectors can be used, those for which $a_i^2 \lambda_i$ is largest should be chosen. However, this can cause the model to not generalize well to unseen data. To avoid this, it is customary to choose the eigenvectors with the largest eigenvalues.

Partial Least Squares (PLS)

- Principal component regression does not address the correlation with the regressand.
- *Partial Least Squares (PLS)* aims to maximize the correlation between regressors and regressand.
- This time a new set of regressors are generated iteratively as linear combinations of regressors, in the first iteration say

$$z_1\mathbf{d}_1 + \ldots + z_M\mathbf{d}_M = \mathbf{D}\mathbf{z}.$$

- Wlog $\|\mathbf{D}\mathbf{z}\| = 1$, since regressors are invariant to scaling.

PLS

- The matrix $\mathbf{D}^T\mathbf{t}\mathbf{t}^T\mathbf{D}$ is a symmetric, positive semidefinite $M \times M$ matrix, since

$$(\mathbf{D}^T\mathbf{t}\mathbf{t}^T\mathbf{D})^T = \mathbf{D}^T\mathbf{t}\mathbf{t}^T\mathbf{D} \text{ and } \mathbf{v}^T\mathbf{D}^T\mathbf{t}\mathbf{t}^T\mathbf{D}\mathbf{v} = (\mathbf{v}^T\mathbf{D}^T\mathbf{t})^2 \geq 0.$$

- It has M non-negative eigenvalues and corresponding orthonormal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_M$ where the eigenvectors are ordered with regards to the corresponding eigenvalues from largest to smallest.
- $\mathbf{z} \in \mathbb{R}^M$, and thus can be expressed as a linear combination of these eigenvectors:

$$\mathbf{z} = \hat{z}_1\mathbf{v}_1 + \ldots + \hat{z}_M\mathbf{v}_M.$$

PLS

- The square of the correlation between $\mathbf{t}$ and the new regressor $\mathbf{Dz}$ is

$$\left(\frac{\mathbf{t}^T\mathbf{Dz}}{\|\mathbf{t}\|_2\|\mathbf{Dz}\|_2}\right)^2 = \frac{1}{\|\mathbf{t}\|_2^2}\mathbf{z}^T\mathbf{D}^T\mathbf{t}\mathbf{t}^T\mathbf{Dz} = \frac{1}{\|\mathbf{t}\|_2^2}\left(\lambda_1\hat{z}_1^2 + \ldots \lambda_M\hat{z}_M^2\right).$$

- This is maximal for $\hat{z}_2 = \ldots = \hat{z}_M = 0$.
- Thus the first new regressor is $\mathbf{t}_1 = \mathbf{Dv}_1/\|\mathbf{Dv}_1\|_2$.

PLS

- Having generated $\mathbf{t}_1$, we calculate

$$\mathbf{D}_1 = \left(\mathbf{I} - \mathbf{t}_1\mathbf{t}_1^T\right)\mathbf{D}.$$

- Note

$$\mathbf{D}_1\mathbf{v}_1 = \left(\mathbf{I} - \mathbf{t}_1\mathbf{t}_1^T\right)\mathbf{D}\mathbf{v}_1 = \|\mathbf{D}\mathbf{v}_1\|_2(1 - \|\mathbf{t}_1\|_2^2)\mathbf{t}_1 = 0.$$

- Let $\mathbf{v}_2$ with $\|\mathbf{v}_2\|_2 = 1$ be the eigenvector corresponding to the largest eigenvalue of $\mathbf{D}_1^T\mathbf{t}\mathbf{t}^T\mathbf{D}_1$. The second new regressor is $\mathbf{t}_2 = \mathbf{D}_1\mathbf{v}_2$ normalized such that $\|\mathbf{t}_2\|_2 = 1$. Again, the correlation is maximal.
- The process continues until $\mathbf{D}_r$ is a null matrix, i.e. its rank is zero.
- We have $\mathrm{rank}\mathbf{D}_j \leq \mathrm{rank}\mathbf{D}_{j-1} - 1$ since a vector $(\mathbf{v}_j)$ which previously was not mapped to zero, now is mapped to zero.
- (PLS can be used for multivariate regression, that is the regressand is a matrix $\mathbf{t}$ of size $N \times q$.)

Regularization

- Recall, we are trying to find suitable coefficients $\mathbf{c}$ minimizing $\|\mathbf{t}-\mathbf{Dc}\|_2^2$ while avoiding unnecessary complexity which leads to over-fitting.
- This is achieved by introducing a *penalty term* $\Omega(\mathbf{c})$ and minimizing

$$\|\mathbf{t}-\mathbf{Dc}\|^2+\lambda\Omega(\mathbf{c}).$$

- The penalty term is also known as *entropy measure*.
- λ controls the trade-off between fitting the data and reducing complexity and has to be optimized itself.
- The choices for $\Omega(\mathbf{c})$ are numerous.

L_0 regularization

L_0 regularization $\Omega(\mathbf{c}) = \|\mathbf{c}\|_0 =$ number of non-zero components.

- Aim is *sparse* vector of coefficients, where many entries are zero.
- If $\mathbf{c}$ is *1-sparse*, we have $\binom{M}{1}$ possibilities.
- If $\mathbf{c}$ is *k-sparse*, $\binom{M}{k}$ possibilities need to be checked.
- We do not not know beforehand, how many components are non-zero. All M possible values for k need to be checked.
- The complexity is

$$\sum_{k=1}^{M} \binom{M}{k} = 2^M - 1.$$

- NP hard problem.
- Related to the *Bayesian Information Criterion* and the *Akaike Information Criterion*.

L_1 regularization

L_1 regularization $\Omega(\mathbf{c}) = \|\mathbf{c}\|_1 = \sum_{m=1}^{M} |c_m|.$

- No unique minimum in the case of perfect multicollinearity, but continuum of minima.
- For example, if two columns $\mathbf{d}_i$ and $\mathbf{d}_j$ are the same and c_i and c_j are nonzero coefficients in the minimal solution, then for any $a \in [0,1]$ replacing c_i by $a(\mathbf{c}_i + \mathbf{c}_j)$ and replacing $\mathbf{c}_j$ by $(1-a)(\mathbf{c}_i + \mathbf{c}_j)$ is also minimal.

Soft thresholding

Soft thresholding

- If the columns of $\mathbf{D}$ are orthonormal, then $\mathbf{D}^T\mathbf{D} = \mathbf{I}$.
- $\mathbf{c}^{\mathrm{OLS}} = \left(\mathbf{D}^T\mathbf{D}\right)^{-1}\mathbf{D}^T\mathbf{t} = \mathbf{D}^T\mathbf{t}$.
- Minimizing $\frac{1}{2}\|\mathbf{t} - \mathbf{D}\mathbf{c}\|_2^2 + \lambda\|\mathbf{c}\|_1$ is equivalent to minimizing

$$-\mathbf{c}^T\mathbf{c}^{\mathrm{OLS}} + \frac{1}{2}\mathbf{c}^T\mathbf{c} + \lambda\|\mathbf{c}\|_1 = \sum_{m=1}^{M} -c_m c_m^{\mathrm{OLS}} + \frac{1}{2}c_m^2 + \lambda|c_m|.$$

- For each m, we minimize

$$-c_m c_m^{\mathrm{OLS}} + \frac{1}{2}c_m^2 + \lambda|c_m| = -c_m c_m^{\mathrm{OLS}} + \frac{1}{2}c_m^2 + \lambda \mathsf{sgn}(c_m)c_m,$$

- c_m and c_m^{OLS} must have the same sign.

Soft thresholding

- Differentiating with respect to c_m and setting to zero, gives

$$c_m = c_m^{\mathrm{OLS}} - \lambda \mathrm{sgn}(c_m^{\mathrm{OLS}}) = \mathrm{sgn}(c_m^{\mathrm{OLS}})(|c_m^{\mathrm{OLS}}| - \lambda).$$

- To ensure that c_m has the same sign as c_m^{OLS}, we set $c_m = 0$, if $|c_m^{\mathrm{OLS}}| < \lambda$. This effectively prunes the corresponding basis function.
- For $\lambda = 0$, we recover the ordinary least square solution.
- If λ is chosen too large, all coefficients are set to zero.

L_2 regularization

L_2 regularization $\Omega(\mathbf{c}) = \|\mathbf{c}\|_2 = \sum_{m=1}^{M} c_m^2$.

- Minimizing

$$\frac{1}{2}\|\mathbf{t} - \mathbf{D}\mathbf{c}\|_2^2 + \frac{1}{2}\lambda\|\mathbf{c}\|_2^2 = \frac{1}{2}\mathbf{t}^T\mathbf{t} - \mathbf{c}^T\mathbf{D}^T\mathbf{t} + \frac{1}{2}\mathbf{c}^T(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})\mathbf{c}.$$

- The gradient is

$$-\mathbf{D}^T\mathbf{t} + (\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})\mathbf{c}.$$

- Setting the gradient to zero,

$$\mathbf{c} = (\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^T\mathbf{t}.$$

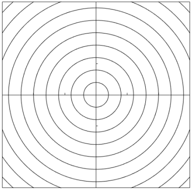
Regularization

In the case $\mathbf{D}^T\mathbf{D} = \mathbf{I}$

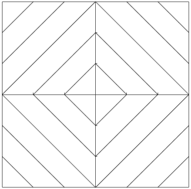
- L_2 regularization reduces each component of the OLS solution by a factor of $(1+\lambda)^{-1}$.
- L_1 regularization moves each component towards zero by λ. If the component is closer to zero than λ, it is set to zero.

L_p norm

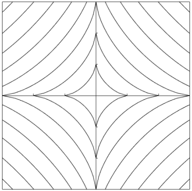
The L_p *norm* is defined as $\|\mathbf{c}\|_p = \left(\sum_{m=1}^{M} |c_m|^p \right)^{1/p}$.



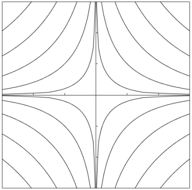
$p = 2$



$p = 1$



$p = 2/3$



$p = 1/3$

Regularization - Penalty

$\Omega(\mathbf{c})$	Regression method
$\|\mathbf{c}\|_0 =$ number of nonzero elements in $\mathbf{c}$	L_0 regularization Bayesian Information Criterion (BIC) Akaike Information Criterion (AIC)
$\|\mathbf{c}\|_1 = \sum_{m=1}^{M} \|c_m\|$	L_1 regularization Least Absolute Shrinkage and Selection Operator (LASSO)
$\|\mathbf{c}\|_2^2 = \sum_{m=1}^{M} c_m^2$	L_2 regularization Ridge regression
$\lambda\|\mathbf{c}\|_1 + \frac{1-\lambda}{2}\|\mathbf{c}\|_2^2$	Elastic net regularization

Bayesian Regression

- Aim to find suitable coefficients $\mathbf{c}$ to satisfy $\mathbf{t} = \mathbf{D}\mathbf{c} + \boldsymbol{\epsilon}$.

- Minimizing $\|\mathbf{t} - \mathbf{D}\mathbf{c}\|^2 = \|\boldsymbol{\epsilon}\|^2$ leads to the noise being modeled and over-fitting.

- Minimizing $\|\mathbf{t} - \mathbf{D}\mathbf{c}\|^2 + \Omega(\mathbf{c})$ "tweaks" the model to favour less complex models.

- Can the "tweaking" be formalized?

Bayes Rule

Bayes Rule: $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

- A and B are events.
- $P(A)$ and $P(B)$ are the probabilities of A and B without regard to each other.
- $P(A|B)$ is the *conditional probability* of A given that B is true. $P(B|A)$ is the conditional probability of B given that A is true.
- Often expressed as $P(A|B) \propto P(B|A)P(A)$ where $\propto$ means that the two sides are proportional to each other.

Bayes

- Bayes rule:
$$p(\mathbf{c}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{c})p(\mathbf{c})}{p(\mathbf{t})}.$$
- That is, we answer the question: What is the probability of the coefficients given the observed data?
- Since the noise is i.i.d. normal with mean 0 and variance σ^2, we can write
$$p(\mathbf{t}|\mathbf{c},\sigma^2) = (2\pi)^{-N/2}\sigma^{-N}\exp\left(-\frac{\|\mathbf{t}-D\mathbf{c}\|}{2\sigma^2}\right).$$

Prior

- We define a *prior distribution* $p(\mathbf{c})$ using all information apart from the data itself quantifying our belief about the coefficients.
- For example, a simple assumption is that each coefficient is a priori normally distributed with mean zero and variance α^{-1},

$$p(\mathbf{c}|\alpha) = \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left(-\frac{\alpha}{2}\mathbf{c}^T\mathbf{c}\right).$$

- α is a *hyperparameter* and known as the *precision* of the distribution. If α becomes very large the distribution becomes peaked at its mean and we have more confidence in the value than if α is small and the width of the distribution large.

Posterior

- After observing N samples of the dependent variable $\mathbf{t}$, the *posterior distribution* is given by

$$p(\mathbf{c}|\mathbf{t}, \mathbf{D}, \alpha, \sigma^2) \propto p(\mathbf{t}|\mathbf{D}, \mathbf{c}, \sigma^2)p(\mathbf{c}|\alpha).$$

- An estimate of $\mathbf{c}$ could be obtained by choosing the $\mathbf{c}$ where the posterior distribution or equivalently its logarithm is maximal,

$$\log p(\mathbf{c}|\mathbf{t}, \mathbf{D}, \alpha, \sigma^2) \propto \log p(\mathbf{t}|\mathbf{D}, \mathbf{c}\, \sigma^2) + \log p(\mathbf{c}|\alpha).$$

- The first term is exactly the log likelihood, which is maximized when $\|\mathbf{t} - \mathbf{D}\mathbf{c}\|^2$ is minimal.
- The second term is $-\dfrac{\alpha}{2}\|\mathbf{c}\|^2 +$ a constant, and can be regarded as the negative of a penalty term.
- In fact, this has re-created the ridge regression.

Bayesian Regression

- More generally let $\boldsymbol{\alpha}$ contain all parameters governing the joint distribution $p(\mathbf{t}, \mathbf{c}|\mathbf{D}, \boldsymbol{\alpha}, \sigma^2)$.
- We remove the dependency on the latent variables by integrating over the coefficients $\mathbf{c}$. Thus averaging over all possible solutions.
- This is called *marginalizing* over $\mathbf{c}$ and the result is the *marginal likelihood*

$$\mathcal{L}(\mathbf{t}|\mathbf{D}, \boldsymbol{\alpha}, \sigma^2) = \int p(\mathbf{t}, \mathbf{c}|\boldsymbol{\alpha}, \sigma^2) d\mathbf{c},$$

  which we aim to maximize.
- The *Maximum-Likelihood Estimate (MLE)* for $\boldsymbol{\alpha}$ are the values which maximize $\mathcal{L}(\mathbf{t}|\mathbf{D}, \boldsymbol{\alpha}, \sigma^2)$.

Expectation-Maximization(EM)

- The *Expectation-Maximization (EM)* algorithm arrives at a maximum iteratively by alternating between two steps.
- First note that for any distribution $q(\mathbf{c})$,

$$\log \mathcal{L}(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2) = \int q(\mathbf{c}) \log \frac{p(\mathbf{t}, \mathbf{c}|\boldsymbol{\alpha}, \sigma^2)}{q(\mathbf{c})} d\mathbf{c} - \int q(\mathbf{c}) \log \frac{p(\mathbf{c}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2)}{q(\mathbf{c})} d\mathbf{c}.$$

- The second term is the Kullback–Leibler divergence between the distributions $q(\mathbf{c})$ and $p(\mathbf{c}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2)$ which is always non-negative and thus the first term is a lower bound for $\log \mathcal{L}(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2)$.
- The maximization is done by maximizing the lower bound by alternating between maximizing with respect to $q(\mathbf{c})$ and with respect to $\boldsymbol{\alpha}$.

EM

- Let $\boldsymbol{\alpha}^{(k)}$ be the current estimate of the maximal $\boldsymbol{\alpha}$.
- The lower bound is maximal wrt. $q(\mathbf{c})$ if $q(\mathbf{c}) = p(\mathbf{c}|\mathbf{t}, \boldsymbol{\alpha}^{(k)}, \sigma^2)$.
- The next step is to find $\boldsymbol{\alpha}^{(k+1)}$ by maximizing

$$\int p(\mathbf{c}|\mathbf{t}, \boldsymbol{\alpha}^{(k)}, \sigma^2) \log \frac{p(\mathbf{t}, \mathbf{c}|\boldsymbol{\alpha}, \sigma^2)}{p(\mathbf{c}|\mathbf{t}, \boldsymbol{\alpha}^{(k)}, \sigma^2)} d\mathbf{c} = \\ \int p(\mathbf{c}|\mathbf{t}, \boldsymbol{\alpha}^{(k)}, \sigma^2) \log p(\mathbf{t}, \mathbf{c}|\boldsymbol{\alpha}, \sigma^2) d\mathbf{c} \\ - \int p(\mathbf{c}|\mathbf{t}, \boldsymbol{\alpha}^{(k)}, \sigma^2) \log p(\mathbf{c}|\mathbf{t}, \boldsymbol{\alpha}^{(k)}, \sigma^2) d\mathbf{c}.$$

- The second term is independent of $\boldsymbol{\alpha}$. Hence only the first term is relevant for the maximization.
- Note: The first term is the *expectation* of the logarithm of the joint probability with respect to the current posterior distribution.

Sparse Bayesian Learning

- As a specific example let the prior distribution of c_i be normal with mean zero and variance α_i^{-1}. That is each weight has its own hyperparameter.
- The multivariate prior is given by

$$p(\mathbf{c}|\boldsymbol{\alpha}) = (2\pi)^{-M/2} \prod_{m=1}^{M} \sqrt{\alpha_m} \exp\left(-\frac{\alpha_m c_m^2}{2}\right).$$

- The multivariate posterior distribution is also normal, since it is derived from the product of two normal distributions.
- It has mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ given by

$$\boldsymbol{\Sigma} = \left(\mathbf{A} + \sigma^{-2}\mathbf{D}^T\mathbf{D}\right)^{-1} \qquad \boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\Sigma}\mathbf{D}^T\mathbf{t},$$

where $\mathbf{A}$ is a diagonal matrix with entries $A_{mm} = \alpha_m$.

Sparse Bayesian Learning

- The logarithm of the marginal likelihood $\mathcal{L}(\mathbf{t}|\mathbf{D}, \boldsymbol{\alpha}, \sigma^2)$ can be calculated analytically,

$$\log \mathcal{L}(\mathbf{t}|\mathbf{D}, \boldsymbol{\alpha}, \sigma^2) = -\frac{1}{2}\left(N \log 2\pi + \log|\mathbf{C}| + \mathbf{t}^T \mathbf{C}^{-1} \mathbf{t}\right),$$

  where $\mathbf{C} = \sigma^2 \mathbf{I} + \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T$.

- $\log \mathcal{L}(\mathbf{t}|\mathbf{D}, \boldsymbol{\alpha}, \sigma^2)$ can be maximized with respect to a single hyperparameter.
- In practice many α_m become infinite during maximization, meaning that the posterior distribution of the corresponding c_m is infinitely peaked at 0 and the corresponding regressor can be removed from the model.

Sparse Bayesian Learning

- The measurement t_n for data $\mathbf{x}_n$ is seen as drawn from a univariate normal distribution with

$$\begin{array}{ll} \text{mean} & m_n = \mathbf{d}(\mathbf{x}_n)^T \boldsymbol{\mu}, \\ \text{variance} & \sigma_n^2 = \sigma^2 + \mathbf{d}(\mathbf{x}_n)^T \Sigma \mathbf{d}(\mathbf{x}_n). \end{array}$$

- If the variance is small, it indicates that at this point the model explains the data well. If the variance is large, the model is not adequate at this point.

RVM

- The *Relevance Vector Machine (RVM)* is an implementation of this technique, where the basis functions are kernel functions centred on $\mathbf{x}_1, \ldots, \mathbf{x}_n$.
- The data points for which α_m remains finite, are known as the *relevant vectors*. The ones necessary to explain the data.