Chapter 6 Regression

- Best Fit to Discrete Data
  - Suppose that n experiments create data set: $(x_1, y_1)\,,\ (x_2, y_2)\,, ..., (x_n, y_n)$.
  - Can we find simple relation $y = \psi(x, a)$ (a is a parameter vector) such that $|y_j - \psi(x_j, a)|\ is$ smallest possible for $j = 1, 2, ..., n$?
  - We call it data fitting, or best-fit. What means by "Best Fit"
  - There are various definitions of "Best-Fit. Most common are two
    * $L^p - fitting$: minimize the L^p error

$$e_p = \sum_{j=1}^{n} |y_j - \psi(x_j, a)|^p$$

    * $L^\infty - fitting$: minimize the L^∞ error

$$e_\infty = \max_{j=1,2,...,n} |y_j - \psi(x_j, a)|$$

  - When $p = 2$, it is called least squares fitting. We shall focus on least squares fitting.
  - Example 2: Linear Regression. We look for a linear function (parameter vector $a = (b, m)$)

$$y = \psi(x, a) = mx + b$$

to fit data $\{(x_1, y_1),\ (x_2, y_2), ..., (x_n, y_n)\}$. The square error is

$$e(m,b) = \sum_{j=1}^{n} |y_j \ - \ \psi(x_j, a)|^2$$
$$= \sum_{j=1}^{n} |y_j \ - m x_j - b\ |^2$$

$*$ To find b, m that would minimizer the above error, we need to have

$$\frac{\partial e(m,b)}{\partial m} = \sum_{j=1}^{n} -2x_j (y_j \ - m x_j - b) \ = 0$$
$$\frac{\partial e(m,b)}{\partial b} = \sum_{j=1}^{n} -2 (y_j \ - m x_j - b) \ = 0$$

$*$ This leads to linear system for m, b

$$m \left(\sum_{j=1}^{n} x_j^2 \right) + b \ \left(\sum_{j=1}^{n} x_j \right) = \sum_{j=1}^{n} x_j y_j$$
$$m \left(\sum_{j=1}^{n} x_j \right) + bn \ = \left(\sum_{j=1}^{n} y_j \right)$$

– In general, the optimal parameter $a = (a_1, a_2, ..., a_m)$ for the square error

$$e\left(a\right) = \sum_{j=1}^{n} \left|y_j \ - \ \psi\left(x_j, a\right)\right|^2$$

is when for all $k - 1, 2, ..., m,$

$$\frac{\partial e\left(a\right)}{\partial a_k} = -2\sum_{j=1}^{n}\left(y_j \ - \ \psi\left(x_j, a\right)\right)\frac{\partial\ \psi\left(x_j, a\right)}{\partial a_k} = 0 \qquad (1)$$

This could be a very complex system for a.

– Example 3 We look for exponential curve $y = \psi\left(x, c, r\right) = ce^{rx}$ to best fit the data. Then

$$\frac{\partial\psi\left(x, c, r\right)}{\partial c} = e^{rx}, \quad \frac{\partial\psi\left(x, c, r\right)}{\partial r} = cxe^{rx}$$

so, with $a_1 = c, \ a_2 = r,$

$$\sum_{j=1}^{n}\left(y_j \ - \ ce^{rx_j}\right)e^{rx_j} = 0$$

$$c\sum_{j=1}^{n}\left(y_j \ - \ ce^{rx_j}\right)x_je^{rx_j} = 0$$

It is impossible to analytically solve c, r. (homework) One alternative is to consider linear regression for the logarithm data: $(x_1, \ln y_1), (x_2, \ln y_2), ..., (x_n, \ln y_n)$.

– Let $y = mx + b$ be the best fitting for above log data. Then (m, b) minimizes

$$\begin{aligned} e(m, b) &= \sum_{j=1}^{n} (\ln y_j \; - mx_j - b)^2 \\ &= \sum_{j=1}^{n} \left(\ln y_j \; - \ln e^{(mx_j + b)}\right)^2 \\ &= \sum_{j=1}^{n} \left(\ln \left[y_j \; e^{-(mx_j + b)}\right]\right)^2 \\ &= \sum_{j=1}^{n} \left(\ln \left[y_j \; e^{-mx_j} e^{-b}\right]\right)^2 \end{aligned}$$

Recall Taylor series for $\ln(x) = \ln(1 - (1 - x)) = x - 1 + O\left((x-1)^2\right)$ .So using linear approximation for $\ln x$,we see

$$\begin{aligned} e\left(m,b\right) &= \sum_{j=1}^{n}\left(\ln\left[y_j\ e^{-mx_j}e^{-b}\right]\right)^2 \\ &\approx \sum_{j=1}^{n}\left(y_j\ e^{-mx_j}e^{-b}-1\right)^2 \\ &= \sum_{j=1}^{n} e^{-2mx_j}e^{-2b}\left(y_j\ -ce^{rx_j}\right)^2, \quad c=e^b,\ r=m, \\ &\sim \sum_{j=1}^{n}\left(y_j\ -ce^{rx_j}\right)^2, \end{aligned}$$

if x_j is bounded.

– Conclusion: linear regression of log data in its first order approximation is equivalent to least square for exponential fitting.

– Example 4 Consider using the following curve to best fit the data:

$$\psi\left(x,a\right)=a_1\phi_1\left(x\right)+a_2\phi_2\left(x\right)+...+a_m\phi_m\left(x\right)$$

where $\phi_1, ..., \phi_m$ are given function. By (1), since $\dfrac{\partial\ \psi(x,a)}{\partial a_i} = \phi_i(x)$,

$$\sum_{j=1}^{n}\left(y_j - \sum_{l=1}^{m} a_l \phi_l(x_j)\right)\phi_i(x_j) = 0,\ i = 1, 2, ..., m$$

or

$$\sum_{j=1}^{n}\left(\sum_{l=1}^{m} \phi_l(x_j)\, a_l\right)\phi_i(x_j) = \sum_{j=1}^{n} y_j \phi_i(x_j).$$

This can also be written as, for $i = 1, 2, ..., m$,

$$\sum_{j=1}^{n}\sum_{l=1}^{m} \phi_i(x_j)\,\phi_l(x_j)\, a_l = \sum_{j=1}^{n} y_j \phi_i(x_j) \tag{2}$$

– Introduce $m \times n$ matrix $\Phi = [\phi_l(x_j)]_{m\times n}$:

$$\Phi = \begin{bmatrix} \phi_1(x_1) & \phi_1(x_2) & \phi_1(x_3) & \cdots & \phi_1(x_n) \\ \phi_2(x_1) & \phi_2(x_2) & \phi_2(x_3) & \cdots & \phi_2(x_n) \\ \phi_3(x_1) & \phi_3(x_2) & \phi_3(x_3) & \cdots & \phi_3(x_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_m(x_1) & \phi_m(x_2) & \phi_m(x_3) & \cdots & \phi_m(x_n) \end{bmatrix}_{m\times n},$$

Then $\Phi\Phi^T$ is $m \times m$

$$\vec{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_m \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

So Equation (2) can be written in matrix equation form

$$\Phi\Phi^T a = \Phi y. \tag{3}$$

– To see (3), we notice that matrix multiplication rules: $B = (b_{ij})_{m\times p}\,,\ C = (c_{ij})_{p\times q}\,,\ D = (d_{ij})_{q\times r}\,,\ then$ $BCD = (e_{ij})_{m\times r}$:

$$e_{ij} = \sum_{l=1}^{q}\sum_{k=1}^{p} b_{ik}c_{kl}d_{lj}$$

Now on the left-hand side of (3), $B = \Phi\ (p = n,\ b_{ij} = \phi_i(x_j)), C = \Phi^T\ (q = m,\ c_{ij} = \phi_j(x_i)), a = D\ (r = 1,\ d_{i,1} = a_i),\ and$

$$e_{i,1} = \sum_{l=1}^{q}\sum_{k=1}^{p} b_{ik}c_{kl}d_{l1} = \sum_{k=1}^{n}\sum_{l=1}^{m} \phi_i(x_k)\,\phi_l(x_k)\,a_l$$

which is the right-hand side of (3), and $\Phi y = (g_{i,1})_{n\times 1}$

$$g_{i,1} = \sum_{k=1}^{n} \phi_i (x_k)\, y_k$$

So (3) is exactly (2).

– For linear regression, $\phi_1 (x) = x,\ \phi_2 (x) = 1$. So $m = 2$, and Φ is $2 \times n$ matrix

$$\Phi = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}$$

and

$$\Phi\Phi^T = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ \cdots & \cdots \\ x_n & 1 \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} x_j^2 & \sum_{j=1}^{n} x_j \\ \sum_{j=1}^{n} x_j & n \end{bmatrix}$$

– So linear regression is to solve $\Phi\Phi^T a = \Phi y$ with $a = (m, b)$.

– Example 5 Polynomial Regression: $\psi = a_0 + a_1 x + a_2 x^2 + ... + a_m x^m$. In this case, $a = (a_0, a_1, ..., a_m)^T,\ \phi_k =$

x^k for $k = 0, 1, ..., m$, and

$$\Phi = \begin{bmatrix} \phi_0(x_1) & \phi_0(x_2) & \phi_0(x_3) & \cdots & \phi_0(x_n) \\ \phi_1(x_1) & \phi_1(x_2) & \phi_1(x_3) & \cdots & \phi_1(x_n) \\ \phi_2(x_1) & \phi_2(x_2) & \phi_2(x_3) & \cdots & \phi_2(x_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_m(x_1) & \phi_m(x_2) & \phi_m(x_3) & \cdots & \phi_m(x_n) \end{bmatrix}_{(m+1)\times n}$$

$$= \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^m & x_2^m & x_3^m & \cdots & x_n^m \end{bmatrix}_{(m+1)\times n}$$

So

$$\Phi\Phi^T = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^m & x_2^m & x_3^m & \cdots & x_n^m \end{bmatrix} \begin{bmatrix} 1 & x_1 & x_1^2 & \vdots & x_1^m \\ 1 & x_2 & x_2^2 & \vdots & x_2^m \\ 1 & x_3 & x_3^2 & \vdots & x_3^m \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ 1 & x_n & x_n^2 & \vdots & x_n^m \end{bmatrix}$$

$$
= \begin{bmatrix}
n & \sum_{j=1}^{n} x_j & \sum_{j=1}^{n} x_j^2 & \cdots & \sum_{j=1}^{n} x_j^m \\
\sum_{j=1}^{n} x_j & \sum_{j=1}^{n} x_j^2 & \sum_{j=1}^{n} x_j^3 & \cdots & \sum_{j=1}^{n} x_j^{m+1} \\
\sum_{j=1}^{n} x_{1j}^2 & \sum_{j=1}^{n} x_j^3 & \sum_{j=1}^{n} x_j^4 & \cdots & \sum_{j=1}^{n} x_j^{m+2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\sum_{j=1}^{n} x_j^m & \sum_{j=1}^{n} x_j^{m+1} & \sum_{j=1}^{n} x_j^{m+2} & \cdots & \sum_{j=1}^{n} x_j^{2m}
\end{bmatrix}_{(m+1)\times(m+1)} = \left[\sum_{j=1}^{n} x_j^{k+l} \right]
$$

– This matrix is invertible for $n > m$ (Exercise 6.8)

- General setting of best-fitting:
  – For linear regression,

$$
\Phi = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}
$$

minimizing

$$e\left(m,b\right)=\sum_{j=1}^{n}\left|y_j\ -\ \psi\left(x_j,a\right)\right|^2=\sum_{j=1}^{n}\left|\ \ y_j-\left(mx_j+\right)b\right|^2$$

$$=\left\|\begin{bmatrix} y_1-(b+mx_1) \\ y_2-(b+mx_2) \\ \vdots \\ y_n-(b+mx_n) \end{bmatrix}\right\|^2=\left\|y-\Phi^T\begin{bmatrix} b \\ m \end{bmatrix}\right\|^2$$

is equivalent to find distance from y to the range of Φ^T

$$dist\left(y,R\left(\Phi^T\right)\right)=\sqrt{\min_{b,m}\left\|y-\Phi^T\begin{bmatrix} b \\ m \end{bmatrix}\right\|^2}$$

– For general regression by $\psi(x,a) = a_1\phi_1(x) + a_2\phi_2(x) + ... + a_m\phi_m(x)$, minimizeing

$$\begin{aligned} e(a) &= \sum_{j=1}^{n} |y_j \ - \ \psi(x_j, a)|^2 \\ &= \sum_{j=1}^{n} \left| y_j \ - \sum_{k=1}^{m} \ a_k\phi_k(x_j) \right|^2 \\ &= \left\| y - \Phi^T a \right\|^2 \end{aligned}$$

is again equivalent to find the distance from y to the range of the $n \times m$ matrix Φ^T

$$dist\left(y, \ R\left(\Phi^T\right)\right) = \left(\min_a \left\| y - \Phi^T a \right\|^2\right)^{1/2}$$

– Range of a matrix is a subspace of R^n. So square best-fitting problem is basically the problem of find distance to a subspace with square norm. What about other norms?

- Section 6.2 Norms in R^n
- A norm on a vector space V is a non-negative mapping/function $x \longmapsto \|x\|$ for any $x \in V$ satisfying

(a) $\|x\| \geq 0$, and $\|x\| = 0$ iff $x = 0$

(b) $\|\lambda x\| = |\lambda| \, \|x\|$

(c) $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)

– Example of norms in R^n : for $x = (x_1, x_2, ..., x_n)$

(i) $\|x\|_\infty = \max_{j=1,2,...,n} |x_j|$ (maximum norm)

(ii) $\|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p}$, $p \geq 1$ (L^p norm or $p - norm$)

$\cdot\ p = 1,\ \|x\|_1 = \sum_{j=1}^{n} |x_j|$

$\cdot\ p = 2,\ \|x\|_2 = \sqrt{\sum_{j=1}^{n} x_j^2}$. This is the familiar distance norm.

- A vector space with a norm is called a normed vector space, or simply normed space.
- Any normed defines a metric (distance): $dist\,(x, y) = \|x - y\|$ .Therefore, it defines a concept of convergence: $x_j \to x$ iff $\|x_j - x\| \to 0$. From there we may define concepts of open sets, boundary of a set, closed set, etc. In other words, it defines a topology on R^n.
- Theorem In R^n, all norms are topologically equivalent, i.e., they define the same convergence. Moreover, for any two norms $\|\|$ and $\|\|'$, there exist two positive constants c_1 and c_2 such that

$$c_1 \|x\| \leq \|x\|' \leq c_2 \|x\| \quad \text{for all } x \in R^n$$

– Proof: Exercise 6.15

- Thus L^p best fit or exponential fit are all equivalent to square fit.
- Section 6.3 Hilbert Space
- Recall a vector space V is a set equipped with addition "$+$" and scalar multiplication "$\cdot$" satisfying 8 properties (vector space axioms)
  - Let $\vec{u}$, $\vec{v}$ and $\vec{w}$ be three vectors in V, λ and δ be two real numbers. Then
    (i) $\vec{u}+\vec{v}=\vec{v}+\vec{u}$
    (ii) $\vec{u}+(\vec{v}+\vec{w})=(\vec{u}+\vec{v})+\vec{w}$
    (iii) $\vec{u}+\vec{0}=\vec{u}$
    (iv) $\vec{u}+(-\vec{u})=\vec{0}$
    (v) $\lambda(\vec{u}+\vec{v})=\lambda\vec{u}+\lambda\vec{v}$
    (vi) $(\lambda+\delta)\vec{u}=\lambda\vec{u}+\delta\vec{u}$
    (vii) $(\lambda\delta)\vec{u}=\lambda(\delta\vec{u})$
    (viii) $1\cdot\vec{u}=\vec{u}$
  - Example of finite dimensional vector space: R^n
  - Example of infinite dimensional vector space: $C^n[0,1]$, $P_n=$polynomials with degree $\leq n$.
- Definition: Consider V be a vector space of finite or infinite dimension. An inner product on V is a symmetric, positive definite bilinear mapping $\langle\cdot,\cdot\rangle : V\times V\to R,\ satisfyibg$
  (a) $\langle x,y\rangle=\langle y,x\rangle$

(b) $\langle x, x \rangle \geq 0$ with equality exactly when $x = 0$

(c) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$

(d) $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle$

- For inner product, we define the norm $\|\cdot\|$ induced by the inner product as

$$\|x\| = \sqrt{\langle x, x \rangle} \tag{4}$$

Exercise: use inequality (5) below to prove (40 is a norm.

- The Cauchy-Schwarz Inequality:

$$|\langle x, y \rangle| \leq \|x\| \, \|y\| \tag{5}$$

Proof: Exercise 6.16. Hint: expand out $\langle x - cy, x - cy \rangle \, , \; c = x / \|y\|$.

- So an inner product induces a normed space, and thus induces the concept of convergence and a topology

- An infinite sequence $\{x_n\}_{n=1}^{\infty}$ in a normed space V is called a Cauchy sequence if

$$\|x_n - x_m\| \longrightarrow 0 \quad \text{as} \quad n, m \longrightarrow \infty$$

- A normed space is called complete if any Cauchy sequence converges.

- Definition: A vector space equipped with an inner product that induces a complete normed space is called a Hilbert Space.

- Orthogonal sequence $x_1, x_2, ... : \langle x_i, x_j \rangle = 0$ if $i \neq j$
- Orthonormal sequence $x_1, x_2, ... :$ it is an orthogonal sequence with unit vector, i.e., $\langle x_i, x_i \rangle = 1$
- Orthonormal basis is an orthonormal set $\{\phi_n\}_{n \in \Omega}$ such that for any vector $f \in V$ can be expressed as

$$f = \sum_{n \in \Omega} c_n \phi_n, \quad \Omega \text{ is a set of index}$$

- A Hilbert space with a countably infinite Orthonormal basis (i.e., Ω is a set of countably many elements) is called a separable Hilbert space. In this case, the above expression becomes

$$f = \sum_{n=1}^{\infty} c_n \phi_n, \quad c_n = \langle f, \phi_n \rangle \text{ is called the n-th coordinate}$$

- Example 7: l^2 consists of all infinite sequences $\{x_n\}_{n=1}^{\infty}$ of real numbers x_n satisfying

$$\sum_{n=1}^{\infty} x_n^2 < \infty$$

– A standard basis in l^2 is

$$\phi_1 = (1, 0, 0, 0, ...)^T$$
$$\phi_2 = (0, 1, 0, 0, ...)^T$$
$$...$$

– almost all properties holds in R^n hold in l^2.

- All separable Hilbert space may be viewed as l^2.
- Example 8: $L^p [a, b] : p = 2$ is Hilbert space, but for $p \neq 2$, it is not.
  – $L^2 [a, b]$ is separable: any square-integrable function is L^2limit of continuous functions which are also limits of polynomials.
  – $1, x, x^2, ...$ form a basis for L^2.
  – Legendre polynomials form an orthogonal baisis for $L^2 [-1, 1]$:

$$P_0 = 1,\ P_1 = x$$
$$(n+1) P_{n+1} = (2n+1) x P_n (x) - n P_{n-1} (x)$$
$$\int_{-1}^{2} P_m (x) P_n (x) dx = \frac{2}{2n+1} \delta_{mn}$$

– $L^2[-\pi,\pi]$ has orthornormal basis

$$\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\sin(nx), \ \frac{1}{\sqrt{\pi}}\cos(nx), \ n = 1, 2, ...$$

Section 6.4 Gram's Theorem on Regression

- Gram's Theorem: Let X be a Hilbert space, and f, $\phi_1, \phi_2, ..., \phi_n$ are in X.Then the best square approximation of f in the form of

$$\psi = c_1\phi_1 + c_2\phi_2 + ... + c_n\phi_n = \sum_{j=1}^{n} c_j\phi_j \qquad (6)$$

occurs when $c_1, c_2, ..., c_n$ solves

$$\sum_{j=1}^{n} c_j \left\langle \phi_i, \phi_j \right\rangle = \langle \phi_i, f\rangle \quad \text{for } i = 1, 2, ..., n \qquad (7)$$

- The matrix form of (7) $is\ AC = F,$ where $A = \left[\left\langle \phi_i, \phi_j \right\rangle\right]$ is a symmetric metrix, $C = (c_1, ..., c_n)^T$, $F = (\langle \phi_1, f\rangle, \langle \phi_2, f\rangle, ..., \langle \phi_n, f\rangle)^T$.
- **Proof**: Let $e(c_1, ..., c_n) = \|\psi - f\|^2 = \langle \psi - f, \psi - f\rangle$ .Note that

$$\frac{\partial(\psi - f)}{\partial c_i} = \frac{\partial \psi}{\partial c_i} = \phi_i$$

Then

$$\frac{\partial e\left(c_1,...,c_n\right)}{\partial c_i}=2\left\langle\frac{\partial\left(\psi-f\right)}{\partial c_i},\psi-f\right\rangle=2\left\langle\phi_i,\psi-f\right\rangle=0$$

or

$$\left\langle\phi_i,\psi\right\rangle=\left\langle\phi_i,f\right\rangle \tag{8}$$

Expanding out equation (8)

$$\left\langle\phi_i,\psi\right\rangle=\left\langle\phi_i,\sum_{j=1}^{n}c_j\phi_j\right\rangle=\sum_{j=1}^{n}c_j\left\langle\phi_i,\phi_j\right\rangle$$

leads to (7)

- Geometrically, let $S=span\left\{\phi_1,\phi_2,...,\phi_n\right\}$, and ψ_0 be the best approximation. $Then\ \|f-\psi_0\|=dist\left(f,S\right)$, and $\left(f-\psi_0\right)\perp S$.To see this, we pick any $\phi\in S$ and consider

$$h\left(t\right)=\left\|f-\left(\psi_0-t\phi\right)\right\|^2=\left\langle f-\psi_0+t\phi,f-\psi_0+t\phi\right\rangle.$$

Since $\left(\psi_0-t\phi\right)\in S$, this function reaches min at $t=0$. Now we write $e=f-\psi_0$,then

$$h\left(t\right)=\left\langle e+t\phi,e+t\phi\right\rangle=\left\langle e,e\right\rangle+2\left\langle\phi,e\right\rangle+t^2\left\langle\phi,\phi\right\rangle.$$

Since it has a min at $t=0$, $h'(0)=0$, i.e.,

$$h'(0)=\langle\phi,e\rangle=0 \implies (f-\psi_0)\perp\phi$$

Corollary: Gram's Theorem can be extended to $n=\infty$. In other words, let X be a Hilbert space, and $f, \phi_1, \phi_2, ...,$ are in X. Assume that

$$\sum_{j=1}^{\infty}\left\|\phi_j\right\|^2<\infty$$

Then the best square approximation of f in the form of

$$\psi=c_1\phi_1+c_2\phi_2+...=\sum_{j=1}^{\infty}c_j\phi_j, \text{ for } \sum_{j=1}^{\infty}c_j^2<\infty$$

occurs when $c_1, c_2, ...$ solves

$$\sum_{j=1}^{\infty}c_j\left\langle\phi_i,\phi_j\right\rangle=\langle\phi_i,f\rangle \quad \text{for } i=1,2,...$$

• **Bessel's Theorem on Regression**: If $\phi_1, \phi_2, ...,$ is an orthogonal sequence, then the best square fit by (6) occurs when

$$c_i = \frac{\langle f, \phi_i \rangle}{\|\phi_i\|^2}$$

and the best approximation is

$$\psi = \frac{\langle f, \phi_1 \rangle}{\|\phi_1\|^2}\phi_1 + \frac{\langle f, \phi_2 \rangle}{\|\phi_2\|^2}\phi_2 + ... + \frac{\langle f, \phi_n \rangle}{\|\phi_n\|^2}\phi_n,$$

and the inequality holds

$$\sum_{i=1}^{\infty} \frac{\langle f, \phi_i \rangle^2}{\|\phi_i\|^2} \leq \|f\|^2 \tag{9}$$

If moreover, $\phi_1, \phi_2, ...,$ is an orthonormal sequence, then

$$\psi = \langle f, \phi_1 \rangle \phi_1 + \langle f, \phi_2 \rangle \phi_2 + ... = \sum_{i=1}^{\infty} \langle f, \phi_i \rangle \phi_i.$$

$$\sum_{i=1}^{\infty} \langle f, \phi_i \rangle^2 \leq \|f\|^2$$

- Proof: Since $\left\langle \phi_i, \phi_j \right\rangle = \delta_{ij}$, the results follow directly from (7).

- Example 10: Find the best square fit in $L^2[-\pi,\pi]$ of $f(t)$ by

$$f_0=\frac{a_0}{2}+\sum_{n=1}^{\infty}\left(a_n\cos nt+b_n\sin nt\right).$$

Sol: According to Example 8,

$$\frac{1}{\sqrt{2\pi}},\frac{1}{\sqrt{\pi}}\sin(nx),\ \frac{1}{\sqrt{\pi}}\cos(nx),\ n=1,2,...$$

form orthonormal basis. So the best fit can be written as

$$f_0=\frac{c_0}{\sqrt{2\pi}}+\sum_{n=1}^{\infty}\left(c_n\frac{\cos nt}{\sqrt{\pi}}+d_n\frac{\sin nt}{\sqrt{\pi}}\right)$$

where

$$c_0=\int_{-\pi}^{\pi}\frac{f(t)}{\sqrt{2\pi}}dt,\ \ \text{so } a_0=\frac{1}{\pi}\int_{-\pi}^{\pi}f(t)\,dt$$

$$c_n=\int_{-\pi}^{\pi}\frac{f(t)\cos nt}{\sqrt{\pi}}dt,\quad \text{so } a_n=\frac{1}{\pi}\int_{-\pi}^{\pi}f(t)\cos nt dt$$

$$d_n = \int_{-\pi}^{\pi} \frac{f(t)\sin nt}{\sqrt{\pi}} dt, \quad \text{so } b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin nt dt$$

This is exactly the Fourier series of f

- In particular, for

$$f(t) = sgn(\sin t)$$

$$a_n = 0,\ b_n = 2\frac{1+(-1)^n}{n\pi}$$

$$f \to \sum_{n=1}^{\infty} 2\left(\frac{1+(-1)^n}{n\pi}\right)\sin nt$$

- Example: Reconsider Example 4: using the following curve to best fit the data (x_i, y_i), $i = 1, 2, ..., n$:

$$\psi(x, a) = c_1\phi_1(x) + c_2\phi_2(x) + ... + c_m\phi_m(x)$$

Sol: Recall that best-fit is to find $a = (c_1, c_2, ..., c_m)$ to minimize

$$\sum (y_i - \psi(x_i, a))^2$$

Let $V = R^n$,

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \ \Phi_k = \begin{bmatrix} \phi_k(x_1) \\ \vdots \\ \phi_k(x_n) \end{bmatrix}$$

$$\Psi = c_1\Phi_1 + c_2\Phi_2 + ... + c_m\Phi_m = \begin{bmatrix} \psi(x_1, a) \\ \vdots \\ \psi(x_n, a) \end{bmatrix}.$$

Then

$$\sum (y_i - \psi(x_i, a))^2 = \|y - \Psi\|$$

The the problem of best fit data (x_i, y_i) by $\psi(x, a)$ is to best-fit of f is by Ψ in Hilbert space V. So we can now use Gram's Theorem,

$$\sum_{j=1}^{m\infty} c_j \langle \Phi_i, \Phi_j \rangle = \langle \Phi_i, f \rangle \quad for\ i = 1, 2, ..., m$$

This is exactly (2).

- Homework: textbook - #6.3, 6.4, 6.8, 6.22, 6.31, 6.34