Applying Hilbert Spaces to the 'Line of Best Fit' Problem

David Suda

Introduction

Although it is possible to study various mathematical applications without explicit use of Hilbert space terminology and techniques, there are great advantages to be gained from Hilbert space formulation. Concepts with which we are familiar with in two- and three- dimensional Euclidean geometry, in particular orthogonality and projections, can also be appropriately extended to infinite-dimensional Hilbert Spaces. One Hilbert Space application which we shall discuss further on today is the application of Hilbert Spaces to the general linear model.

The properties which identify a Hilbert Space are its completeness and its inner product space properties. With reference to Cauchy Sequences, we hence define the property of completeness in the Hilbert Space context as follows.

Definition 1.1 *A* Hilbert Space $\mathcal{H}$ *is an inner-product space which is complete, i.e., an inner-product space in which every Cauchy Sequence* $\{x_n\}$ *converges in norm to some element* $x \in \mathcal{H}$

The Projection Theorem

A powerful result of Hilbert Spaces is the *Projection Theorem*. The Projection Theorem can be of use to us when due to incomplete knowledge of the basis of a Hilbert space (as in infinite-dimensional Hilbert space), or when we do not want to use the complete basis of the Hilbert space (as in parsimony of variables in many statistical applications), we intend to find the best possible approximation of an element of the Hilbert space. To illustrate this we shall take the following example in $\mathbb{R}^3$. Suppose we are given a vector $\mathbf{y} = \left(\frac{1}{4}, \frac{1}{4}, 1\right)'$. Suppose also that we shall limit ourselves to the use of the two linearly independent vectors $\mathbf{x_1} = \left(1, 0, \frac{1}{4}\right)'$ and $\mathbf{x_2} = (0, 1, \frac{1}{4})'$. Our problem is to find the linear combination $\hat{\mathbf{y}} = \alpha_1\mathbf{x_1} + \alpha_2\mathbf{x_2}$. Also, since we want to minimise the distance between $\mathbf{y}$ and $\hat{\mathbf{y}}$ as much as possible, we require the vector $\mathbf{y} - \alpha_1\mathbf{x_1} - \alpha_2\mathbf{x_2}$ to be orthogonal to the plane of $\mathbf{x_1}$ and $\mathbf{x_2}$. Hence the orthogonality condition may be stated as:

$$< \mathbf{y} - \alpha_1\mathbf{x_1} - \alpha_2\mathbf{x_2}, \mathbf{x_i} >= 0; i = 1, 2$$

or equivalently:

$$\alpha_1 < \mathbf{x_1}, \mathbf{x_1} > + \alpha_2 < \mathbf{x_1}, \mathbf{x_2} >=< \mathbf{y}, \mathbf{x_1} >$$

$$\alpha_1 < \mathbf{x_1}, \mathbf{x_2} > + \alpha_2 < \mathbf{x_2}, \mathbf{x_2} >=< \mathbf{y}, \mathbf{x_2} >$$

which, for the specified $\mathbf{x_1}$, $\mathbf{x_2}$ and $\mathbf{y}$, these equations become:

$$\frac{17}{16}\alpha_1 + \frac{1}{16}\alpha_2 = \frac{1}{2}$$

$$\frac{1}{16}\alpha_1 + \frac{17}{16}\alpha_2 = \frac{1}{2}$$

from which we deduce that $\alpha_1 = \alpha_2 = \frac{4}{9}$ and $\hat{\mathbf{y}} = \left(\frac{4}{9}, \frac{4}{9}, \frac{2}{9}\right)'$.

Such an illustration is a clear justification of the importance of the Projection Theorem, which we shall now state.

Theorem 1.1 (The Projection Theorem) *If $\mathcal{M}$ is a closed subspace of the Hilbert space $\mathcal{H}$ and $x \in \mathcal{H}$, then there is a unique element $\hat{x} \in \mathcal{M}$ such that $\|x - \hat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|$. Also $\hat{x} \in \mathcal{M}$ and $\|x - \hat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|$ if and only if $\hat{x} \in \mathcal{M}$ and $(x - \hat{x}) \in \mathcal{M}^{\perp}$.*

A natural consequence of this theorem is the projection operator, which is the mapping of an element from the Hilbert space $\mathcal{H}$ to a unique element in the subspace $\mathcal{M}$. This projection operator is defined as follows:

Definition 1.2 *Let $\mathcal{M}$ be a closed subspace of a Hilbert Space $\mathcal{H}$. The operator $P_{\mathcal{M}}$ on $\mathcal{H}$, defined by $P_{\mathcal{M}}x = \hat{x}$ if $\hat{x} \in \mathcal{M}$ and $(x - \hat{x}) \in \mathcal{M}^{\perp}$ is called the* orthogonal projection operator onto $\mathcal{M}$, *or simply the* projection operator, *on $\mathcal{M}$. Also, the vector $\hat{x}$ is called the projection of x onto $\mathcal{M}$.*

An immediate result from this definition is that the mapping $(I - P_{\mathcal{M}})$ projects an element of $\mathcal{H}$ onto $\mathcal{M}^{\perp}$. The projection operator also has the following properties:

1. $P_{\mathcal{M}}(\alpha x + \beta y) = \alpha P_{\mathcal{M}}x + \beta P_{\mathcal{M}}y$
2. $\|x\|^2 = \|P_{\mathcal{M}}x\|^2 + \|(I - P_{\mathcal{M}})\|^2$
3. Each $x \in \mathcal{H}$ has a unique representation $x = P_{\mathcal{M}}x + (I - P_{\mathcal{M}})X$
4. $P_{\mathcal{M}}x_n \to P_{\mathcal{M}}x$ if $\|x_n - x\| \to 0$
5. $x \in \mathcal{M}$ if and only if $P_{\mathcal{M}}x = x$
6. $x \in \mathcal{M}^{\perp}$ if and only if $P_{\mathcal{M}}x = 0$
7. $\mathcal{M}_1 \subseteq \mathcal{M}_2$ if and only if $P_{\mathcal{M}1}P_{\mathcal{M}2}x = P_{\mathcal{M}1}x, \forall x \in \mathcal{H}$

Note that the seventh property makes the projection operator idempotent. Another result we can deduce from the projection theorem are the *prediction equations.* Given a Hilbert Space $\mathcal{H}$ and a closed subspace $\mathcal{M}$, the unique element $\hat{x} \in \mathcal{M}$ closest to an element x in $\mathcal{H}$ satisfies $< x - \hat{x}, y >= 0, \forall y \in \mathcal{M}$

Projection in $\mathbb{R}^n$

Before going on to show the importance of the projection theorem in problems involving approximation, such as the 'line of best fit' problem, we shall go on to define the following terminology.

Definition 1.3 *The* closed span $\overline{sp}\{x_t, t \in T\}$ *of any subset* $\{x_t, t \in T\}$ *of a Hilbert Space* $\mathcal{H}$ *is defined to be the smallest closed subspace of* $\mathcal{H}$ *which contains each element* $x_t, t \in T$.

Definition 1.4 *A set* $\{e_t, t \in T\}$ *of elements of an inner product space is said to be orthogonal if for every* $s, t \in T$ *:*

$$< e_s, e_t > = \begin{cases} 1, & s = t \\ 0, & s \neq t \end{cases}$$

By Gram-Schmidt orthogonalisation, we can express any element in $\mathbb{R}^n$ by means of an orthogonal set of vectors. In $\mathbb{R}^n$, if $\mathcal{M} = \overline{sp}\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ where $\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ is an orthonormal subset of $\mathcal{M}$ and $m < n$, then there is an orthonormal subset $\{\mathbf{e}_{m+1}, \ldots, \mathbf{e}_n\}$ such that the orthogonal subspace $\mathcal{M}^{\perp} = \overline{sp}\{\mathbf{e}_{m+1}, \ldots, \mathbf{e}_n\}$. In this case:

1. $P_{\mathcal{M}}\mathbf{x} = \sum_{i=m+1}^{n} < \mathbf{x}, \mathbf{e}_i > \mathbf{e}_i$
2. $(I - P_{\mathcal{M}})\mathbf{x} = \sum_{i=m+1}^{n} < \mathbf{x}, \mathbf{e}_i > \mathbf{e}_i$

The following theorem enables us to compute $P_{\mathcal{M}}\mathbf{x}$ directly from any specified set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ then:

$$P_{\mathcal{M}}\mathbf{x} = \mathbf{X}\beta$$

where X is the $n \times n$ matrix whose j^{th} column is $\mathbf{x}_j$ and :

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{x}$$

Proof Since $P_{\mathcal{M}}\mathbf{x} \in \mathcal{M}$ we can write :

$$P_{\mathcal{M}}\mathbf{x} = \sum_{i=1}^{m} \beta_i \mathbf{x}_i = \mathbf{X}\beta$$

for some $\beta \in \mathbb{R}^m$. The prediction equations are, in this case, equivalent to:

$$< \mathbf{X}\beta, \mathbf{x}_j > = < \mathbf{x}, \mathbf{x}_j >, j = 1, \ldots, m$$

and in matrix form these equations can be written as:

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{x}$$

The existence of at least one solution for β is guaranteed by the existence of the projection $P_{\mathcal{M}}\mathbf{x}$. The fact that $\mathbf{X}\beta$ is the same for all solutions is guaranteed by the uniqueness of $P_{\mathcal{M}}\mathbf{x}$. The last statement of the theorem trivially follows. ∎

This equation need not necessarily have a unique solution for β, nut this doesn't violate the uniqueness property of the projection theorem since $X\beta$ will be the same for all solutions. There is exactly one solution if and only if $X'X$ is non-singular, and in this case:

$$P_{\mathcal{M}}\mathbf{x} = X(X'X)^{-1}X'\mathbf{x}$$

Hence, the projection mapping in this case will the be projection matrix $X(X'X)^{-1}X'$. Note that one of the properties of this matrix is that it is idempotent, Our final step is now that of applying this theory to linear regression and the general linear model.

Linear Regression and the general linear Model

Consider the problem of finding the best straight line:

$$y = \theta_1 x + \theta_2$$

or equivalently the best values $\hat{\theta}_1, \hat{\theta}_1$ for $\hat{\theta}_1, \hat{\theta}_2 \in \mathbb{R}$ to fit a given set of data points $(z_i, y_i), i = 1 \ldots, n$. In least squares regression, the best estimates $\hat{\theta}_1, \hat{\theta}_2$ are defined to be the values of $\hat{\theta}_1$ and $\hat{\theta}_2$ which minimise the sum:

$$S(\theta_1, \theta_2) = \sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_2)^2$$

of squared deviations of the observations y_i from the fitted values $\theta_1 x + \theta_2$. This problem reduces to that of computing a projection in $\mathbb{R}^n$ as is easily seen by writing $S(\theta_1, \theta_2)$ in the equivalent form:

$$S(\theta_1, \theta_2) = \|\mathbf{y} - \theta_1 \mathbf{x} - \theta_2 \mathbf{1}\|^2$$

where $\mathbf{x} = (x_1, \ldots, x_n)'$, $\mathbf{1} = (1, \ldots, 1)$ and $\mathbf{y} = (y_1, \ldots, y_n)'$. By the projection theorem there is a unique vector $P_{\mathcal{M}}\mathbf{y}$ of the form $\hat{\theta}_1 \mathbf{x} + \hat{\theta}_2 \mathbf{1}$ which minimises $S(\theta_1, \theta_2)$, where $\mathcal{M} = \overline{sp}\{\mathbf{x}, \mathbf{1}\}$. Defining X to be the $n \times 2$ matrix $X = [\mathbf{x}, \mathbf{1}]$ and $\hat{\theta}$ to be the column vector $\hat{\theta} = (\hat{\theta_1}, \hat{\theta_2})'$ we deduce from the previous theorem that:

$$P_{\mathcal{M}}\mathbf{y} = \mathbf{X}\hat{\theta}$$

where :

$$\mathbf{X}'\mathbf{X}\hat{\theta} = \mathbf{X}'\mathbf{y}$$

There is a unique solution of $\hat{\theta}$ if and only if $X'X$ is non-singular. In this case:

$$\hat{\theta} = (X'X)^{-1}X'\mathbf{y}$$

If $X'X$ is singular there are infinitely many solutions of $\hat{\theta}$, however by the uniqueness of $P_{\mathcal{M}}\mathbf{y}$, $\mathbf{X}\hat{\theta}$ is the same for all of them.
This argument similarly applies to the least squares estimation for the general

linear model. The general problem is as follows. Given a set of data points $(x_i^{(1)}, \ldots, x_i^{(m)}, y_i), i = 1, \ldots .n; m \leq n$, we are required to find a value $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)'$ of $\theta = (\theta_1, \ldots, \theta_m)'$ which minimises:

$$S(\theta) = \sum_{i=1}^{n} (y_i - \theta_1 x_i^{(1)} - \cdots - \theta_m x_i^{(m)})^2 = \|\mathbf{y} - \theta_1 \mathbf{x}^{(1)} - \cdots - \theta_m \mathbf{x}^m\|^2$$

where $\mathbf{x}^{(j)} = (x_1^{(j)}, \ldots, x_n^{(j)}), j = 1, \ldots, m, \mathbf{y} = (y_1, \ldots, y_n)'$. By the projection theorem, there is a unique vector $P_{\mathcal{M}}\mathbf{y}$ of the form $\hat{\theta}_1 \mathbf{x}^{(1)} + \cdots + \hat{\theta}_m \mathbf{x}^{(m)}$ which minimises $S(\theta)$, where $\mathcal{M} = \overline{sp}\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$.

Defining X to be the $n \times m$ matrix $X = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}]$ and $\hat{\theta}$ to be the column vector $\hat{\theta} = (\theta_1, \ldots, \theta_m)'$, we deduce that:

$$P_{\mathcal{M}}\mathbf{y} = X\hat{\theta}$$

where:

$$X'X\hat{\theta} = X'\mathbf{y}$$

As in the simple linear regression, $\hat{\theta}$ is uniquely defined if and only if $X'X$ is non-singular, in which case $\hat{\theta} = (X'X)^{-1}X'\mathbf{y}$. If $X'X$ is singular, then there are infinitely many solutions for $\hat{\theta}$ but $X\hat{\theta}$ is the same for all of them.

Example

This is a simple illustration of the theory that has been constructed in the previous chapters. Let us fit a quadratic function of the form:

$$y = \theta_1 x^2 + \theta_2 x + \theta_3$$

to the data:

x	0	1	2	3	4
y	1	0	3	5	8

The matrix X for this problem is:

$$X = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \\ 16 & 4 & 1 \end{pmatrix}$$

giving:

$$(X'X)^{-1} = \frac{1}{40} \begin{pmatrix} 10 & -40 & 20 \\ -40 & 174 & -108 \\ 20 & -108 & 124 \end{pmatrix}$$

The least squares estimate $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)'$, is therefore unique and is found to be:

$$\hat{\theta} = (0.5, -0.1, 0.6)'$$

from $\hat{\theta} = (X'X)^{-1}X'\mathbf{y}$.

The vector of fitted values $X\hat{\theta} = P_{\mathcal{M}}\mathbf{y}$ is given by:

$$X\hat{\theta} = (0.6, 1, 2.4, 4.8, 8.2)'$$

as compared with the vector of observations:

$$\mathbf{y} = (1, 0, 3, 5, 8)'$$