

Least-squares methods

February 15, 2006

1 Introduction

When the data and model manifolds are linear (or locally linear) and the prior, measurement, and theory uncertainties all involve Gaussian distributions (or equivalently we use minimum information pdfs while assuming fixed first and second moments), then least-squares methods are appropriate. The biggest drawback of least-squares methods is their over-emphasis on any “outliers” in the data. The presence of outliers indicates that Gaussian assumptions on your measurements are basically incorrect. These are better handled with generalized Gaussian models with longer tails ($p < 2$).

2 Mathematics of Linear spaces

Because least-squares methods deal with linear spaces, we give a preface regarding the math of linear spaces. The biggest difference with our approach over what you’ve seen before in ECEN671 (or equivalent) is the explicit use of the dual space.

In what follows we will use the notation that \mathbb{D} is the linear data space and \mathbb{M} is the linear model space. An general linear space is denoted \mathbb{V} . The development in these notes differs from the book by Tarantola in that we allow the linear space to be defined over the complex field. Let \mathbb{V} be an n -dimensional linear space with vectors denoted $\mathbf{u}, \mathbf{v}, \mathbf{w}, \dots$. Select a basis over this space $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ and then any element of the space can be written $\mathbf{v} = \sum_i v^i \mathbf{e}_i$. Typically, we will not include the summation sign and write this as $\mathbf{v} = v^i \mathbf{e}_i$. When a variable is repeated in an upper and lower index we will assume a sum is implied. The sum of two elements and the scalar product can be defined in terms of components as:

$$(\mathbf{u} + \mathbf{v})^i = u^i + v^i \quad (\lambda \mathbf{v})^i = \lambda v^i.$$

The *dual* of \mathbb{V} is denoted \mathbb{V}^* and consists of all linear functionals, or forms, over \mathbb{V} . In other words, \mathbb{V}^* is the space of all linear functions mapping \mathbb{V} into the complex-field, \mathbb{C} . If $\boldsymbol{\omega}$ is an element of \mathbb{V}^* , the mapping it implies is

$$\mathbf{v} \mapsto \lambda = \langle \mathbf{v}, \boldsymbol{\omega} \rangle.$$

Written in more traditional “function” form we say that $\boldsymbol{\omega}(\mathbf{v}) = \lambda$. We use the opposite order of the term in braces so that it conforms to your usual understanding of the inner product. The dual space is a linear vector space. The sum of two linear forms and the scalar product of a linear form are given by (notice the conjugate in the dual space).

$$\langle \mathbf{v}, (\boldsymbol{\omega} + \boldsymbol{\nu}) \rangle = \langle \mathbf{v}, \boldsymbol{\omega} \rangle + \langle \mathbf{v}, \boldsymbol{\nu} \rangle \quad \langle \mathbf{v}, \alpha \boldsymbol{\omega} \rangle = \alpha^* \langle \mathbf{v}, \boldsymbol{\omega} \rangle.$$

Because \mathbb{V}^* is a linear vector space, we can find a basis $\{\boldsymbol{\epsilon}^1, \dots, \boldsymbol{\epsilon}^n\}$, and represent the vector $\boldsymbol{\omega} = \omega_i \boldsymbol{\epsilon}^i$. Notice that both the components of the vector space and the components of the dual space are both represented by n complex numbers. The dual space, however, allows us to distinguish between “column” vectors and “row” vectors. Often the dual space is confused with the vector space. We will use lower-indices on the components of the dual space and upper-indices on the components of the vector space. Although not necessary, it is convenient and typical to choose the bases to be bi-orthogonal (or mutually-dual) so that

$$\langle \mathbf{e}_j, \boldsymbol{\epsilon}^i \rangle = \delta_j^i$$

which is 1 when $i = j$ and zero otherwise. With this representation, the dual-product can be written as

$$\langle \mathbf{v}, \boldsymbol{\omega} \rangle = \langle v^i \mathbf{e}_i, \omega_j \mathbf{e}^j \rangle = w_j^* v^i \delta_i^j = w_i^* v^i.$$

which should be a familiar computational-device. Thus, if we represented the components of \mathbf{v} and the the components of $\boldsymbol{\omega}$ as traditional column vectors we could write

$$\langle \mathbf{v}, \boldsymbol{\omega} \rangle = \boldsymbol{\omega}^H \mathbf{v} = w_i^* v^i.$$

If a vector is thought of as an arrow from the origin to a point in space, then an element of the dual space can be thought of as parallel planes (the normal vector passing through the plane is in fact the “vector” isomorphic to the dual-space element). But, the object in the dual space **is** the collection of planes. The inner product can be envisioned as counting how many planes a vector “pierces”. Every vector-space has a dual space.

3 Linear Operators

A linear operator maps from one vector space to another vector space. Because we now have two categories of vector spaces, we have 4 kinds of linear operators that map from one vector space to another. Suppose we have two vector spaces, \mathbb{V} and \mathbb{W} and their dual spaces \mathbb{V}^* and \mathbb{W}^* . We can define 4 categories of linear operators. Because we use the convention that repeated indices where one is upper and one is lower imply a summation, these four categories of linear operator all have slightly different index notations:

Category	Mapping	Index Notation
I	$A : \mathbb{V} \mapsto \mathbb{W}$	$w^i = A^i_j v^j$
II	$A : \mathbb{V} \mapsto \mathbb{W}^*$	$\omega_i = A_{ij} v^j$
III	$A : \mathbb{V}^* \mapsto \mathbb{W}$	$w^i = A^{ij} \nu_j$
IV	$A : \mathbb{V}^* \mapsto \mathbb{W}^*$	$\omega_i = A_i^j \nu_j$

Notice that the variable we are summing over is shifted to the “left-most” position in the corresponding index notation. All of these categories of linear operators can be used in the course of solving problems.

4 (Hermitian) Transpose of a Linear Operator

With the dual space defined in this way we can define an additional operator for every linear operator between two vector spaces. Suppose \mathbf{G} is a linear operator mapping \mathbb{M} to \mathbb{D} given formally by

$$\mathbf{d} = \mathbf{G}\mathbf{m}.$$

In terms of components of \mathbf{d} and \mathbf{m} , this operator can be written using the implicit sum convention as

$$d^i = G^i_j m^j.$$

The *hermitian-transpose*, or just transpose, of \mathbf{G} denoted \mathbf{G}^H is a linear operator mapping \mathbb{D}^* to \mathbb{M}^* . It is defined by the condition that for any $\boldsymbol{\delta} \in \mathbb{D}^*$ and any $\mathbf{m} \in \mathbb{M}$

$$\langle \mathbf{G}\mathbf{m}, \boldsymbol{\delta} \rangle_D = \langle \mathbf{m}, \mathbf{G}^H \boldsymbol{\delta} \rangle_M.$$

In terms of components

$$G^i_j m^j \delta_i = (\mathbf{G}\mathbf{m})^i \delta_i^* = m^j (G^H \boldsymbol{\delta})_j^* = m^j \left[(G^h)_j^i \right]^* \delta_i^*.$$

Since this must be true for all m^j and δ_i we can see that the representation of \mathbf{G}^h is

$$(G^H)_j^i = (G^i_j)^*$$

which in terms of a matrix of number is the standard, Hermitian-transpose. For notational convenience we will use

$$(\mathbf{G}^H)_i^j \equiv G_i^{*j} = G_i^{j*}.$$

In the special case where a linear operator, \mathbf{S} , is considered that maps a linear space \mathbb{V} into its dual space \mathbb{V}^* , then by definition the hermitian transpose also maps \mathbb{V} into \mathbb{V}^* (the dual-space of a dual-space is the original space). If, in this case, $\mathbf{S} = \mathbf{S}^H$, and the operator is called Hermitian symmetric.

5 Scalar Product

Notice the duality-product is similar to an inner product, except the vectors come from different spaces. We will define the inner product by first describing a weighting operator, \mathbf{W} , as a linear, symmetric and positive-definite mapping from a vector space, \mathbb{V} to its dual \mathbb{V}^* . A positive-definite mapping is one that satisfies $\langle \mathbf{v}, \mathbf{W}\mathbf{v} \rangle > 0$, for all $\mathbf{v} \neq 0$. The scalar product between two vectors in \mathbb{V} can be defined as

$$(\mathbf{v}, \mathbf{u}) = \langle \mathbf{v}, \mathbf{W}\mathbf{u} \rangle = W_{ij}^* u^{*j} v^i = W_{ji} u^{*j} v^i.$$

The last equality is possible because \mathbf{W} is hermitian symmetric.

Notice that

$$(\mathbf{u}, \mathbf{v}) = \langle \mathbf{u}, \mathbf{W}\mathbf{v} \rangle = W_{ij}^* v^{*j} u^i = W_{ji}^* v^{*i} u^j = (\mathbf{v}, \mathbf{u})^*$$

so the scalar product satisfies the usual rules of an inner product. The norm of a vector can then be described as

$$\|\mathbf{v}\| = (\mathbf{v}, \mathbf{v}) = \langle \mathbf{v}, \mathbf{W}\mathbf{v} \rangle = W_{ij}^* v^{*j} v^i.$$

For every weighting operator, \mathbf{W} , there is an inverse operator called a covariance operator, \mathbf{C} , that maps from the dual space to the vector space so that $\mathbf{W}\mathbf{C} = \mathbf{I} = \mathbf{C}\mathbf{W}$.

$$\begin{aligned} W_{ij} C^{jk} &= \delta_i^k \\ C^{ij} W_{jk} &= \delta_k^i. \end{aligned}$$

Sometimes, it is more convenient to use the covariance operator instead of the weighting operator, and vice-versa.

6 Adjoint

The adjoint is defined analogously to the transpose operator, except the scalar (inner) product is used instead of the duality product. Specifically, suppose \mathbf{G} is an operator from \mathbb{M} to \mathbb{D} . Then the adjoint, \mathbf{G}^A , is an operator from \mathbb{D} to \mathbb{M} such that for all \mathbf{d} and \mathbf{m}

$$(\mathbf{G}\mathbf{m}, \mathbf{d})_D = (\mathbf{m}, \mathbf{G}^A\mathbf{d})_M.$$

The components of \mathbf{G}^A can be determined by expanding the definition out in components and noting that the relationship has to hold for all \mathbf{d} and all \mathbf{m} .

$$\begin{aligned} (\mathbf{G}\mathbf{m})^i W_{ji}^D d^{j*} &= m^i (\mathbf{G}^A\mathbf{d})^{*j} W_{ji}^M \\ G_k^i m^k W_{ji}^D d^{j*} &= m^i (\mathbf{G}^A)_k^{*j} d^{k*} W_{ji}^M \\ G_i^k m^i W_{jk}^D d^{j*} &= m^i (\mathbf{G}^A)_j^{k*} d^{j*} W_{ki}^M. \end{aligned}$$

If this is going to be valid for all m^i and all d^j , then

$$\begin{aligned} G_i^{k*} W_{kj}^D &= (\mathbf{G}^A)_j^k W_{ik}^M \\ C_M^{mi} G_i^{k*} W_{kj}^D &= (\mathbf{G}^A)_j^k \delta_k^m \\ (\mathbf{G}^A)_j^m &= C_M^{mi} G_i^{k*} W_{kj}^D. \end{aligned}$$

In matrix notation

$$\mathbf{G}^A = \mathbf{C}_M \mathbf{G}^H \mathbf{C}_D^{-1},$$

Thus, we see the adjoint as a composition of three operators: $\mathbf{W}_D \equiv \mathbf{C}_D^{-1}$ that takes a vector from \mathbb{D} to \mathbb{D}^* , the transpose operator, \mathbf{G}^H , that takes the vector from \mathbb{D}^* to \mathbb{M}^* and finally, a covariance operator, \mathbf{C}_M that takes a vector from \mathbb{M}^* to \mathbb{M} .

7 Norms, distances and metric spaces

Notice that if we have defined the inner product using the given weighting, then

$$\|\mathbf{v}\|^2 = (\mathbf{v}, \mathbf{v}) = \langle \mathbf{v}, \mathbf{W}\mathbf{v} \rangle = W_{ij} v^i v^{j*}.$$

Therefore, the distance between two points is

$$ds^2 = \|d\mathbf{x}\|^2 = W_{ij} dx^i dx^{j*},$$

which fits the general form of a metric-tensor, g_{ij} , defined so that

$$ds^2 = g_{ij} dx^i dx^{j*}.$$

In non-linear spaces, the metric tensor defines how distances are calculated locally so that g_{ij} can be a function of position. In the linear examples used here, we are only considering cases where the metric tensor is constant so the space is linear (but not Euclidean so there is still a difference between the space and its dual). In essence we are using the weighting function, \mathbf{W} , to distinguish (via the map) between elements of the vector space, \mathbb{V} and its dual \mathbb{V}^* . The use of this non-identity weighting function requires that we treat contravariant, \mathbf{v} , and covariant, $\boldsymbol{\omega}$, vectors differently.

8 Least-squares Problems

If (1) homogenous density functions in both the model and data are constant, (2) the prior probability density function is Gaussian, and (3) the noise on both the theory and the data is Gaussian, then the posterior pdf (inverse problem solution) can be written

$$\sigma_M(\mathbf{m}) = k \exp[-S(\mathbf{m})],$$

where

$$2S(\mathbf{m}) = \|\mathbf{g}(\mathbf{m}) - \mathbf{d}_{\text{obs}}\|_D^2 + \|\mathbf{m} - \mathbf{m}_{\text{prior}}\|_M^2$$

where the norm is the scalar-product-norm and so includes the covariance operators \mathbf{C}_D and \mathbf{C}_M . Thus, written out more fully (and shortening $\mathbf{d}_{\text{obs}} \equiv \mathbf{d}$, $\mathbf{m}_{\text{prior}} \equiv \mathbf{m}_0$).

$$2S(\mathbf{m}) = (\mathbf{g}(\mathbf{m}) - \mathbf{d})^H \mathbf{C}_D^{-1} (\mathbf{g}(\mathbf{m}) - \mathbf{d}) + (\mathbf{m} - \mathbf{m}_0)^H \mathbf{C}_M^{-1} (\mathbf{m} - \mathbf{m}_0).$$

We can account for “noise on the theory” simply using $\mathbf{C}_D \mapsto \mathbf{C}_D + \mathbf{C}_T$ so that we can forget there are two sources of uncertainty.

We are implicitly assuming in the Gaussian assumption that either the data and model parameters are real, or if they are complex-valued, then the covariance operator

$$\mathbf{C}_Z = \frac{1}{2} E[\mathbf{z}\mathbf{z}^H]$$

is enough to completely describe the uncertainty on the data. Without loss of generality, imaging a zero-mean length N complex-valued vector $\mathbf{z} = \mathbf{z}_R + j\mathbf{z}_I$. We can always treat this as two real-valued vectors stacked on top of each other to produce a length $2N$ vector \mathbf{p} with

$$\mathbf{p} = \begin{bmatrix} \mathbf{z}_R \\ \mathbf{z}_I \end{bmatrix}.$$

Then

$$\mathbf{C}_P = E[\mathbf{p}\mathbf{p}^T] = \begin{bmatrix} E[\mathbf{z}_R\mathbf{z}_R^T] & E[\mathbf{z}_R\mathbf{z}_I^T] \\ E[\mathbf{z}_I\mathbf{z}_R^T] & E[\mathbf{z}_I\mathbf{z}_I^T] \end{bmatrix} \equiv \begin{bmatrix} C_{RR} & C_{RI} \\ C_{IR} & C_{II} \end{bmatrix}$$

Notice that,

$$2\mathbf{C}_Z = E[\mathbf{z}\mathbf{z}^H] = E[(\mathbf{z}_R + j\mathbf{z}_I)(\mathbf{z}_R^T - j\mathbf{z}_I^T)] = C_{RR} + C_{II} + j(C_{IR} - C_{RI}).$$

To specify a general covariance matrix \mathbf{C}_P we would also need to specify a pseudo-covariance matrix:

$$2\mathbf{J}_Z = E[\mathbf{z}\mathbf{z}^T] = E[(\mathbf{z}_R + j\mathbf{z}_I)(\mathbf{z}_R^T + j\mathbf{z}_I^T)] = C_{RR} - C_{II} + j(C_{IR} + C_{RI}).$$

Then, the components of the general covariance matrix in terms of the covariance and pseudo-covariance are found using the fact that

$$\begin{aligned} \mathbf{C}_Z + \mathbf{J}_Z &= C_{RR} + jC_{IR} \\ \mathbf{C}_Z - \mathbf{J}_Z &= C_{II} - jC_{RI}. \end{aligned}$$

A complex-valued random vector is called *proper* if $\mathbf{J}_Z = 0$. In this case, \mathbf{C}_Z , is enough to completely describe the covariance matrix because $C_{RR} = C_{II}$ and $C_{IR} = -C_{RI}$ ($= -C_{IR}^T$). For proper complex-valued random vectors, we can write a Gaussian form for the pdf using hermitian transpose instead of transpose. We assume proper complex-valued random vectors or full real-valued random vectors throughout.

Notice that maximizing the posterior pdf is equivalent to minimizing the squared-error between the observed data and the model-predicted data and the squared-error between the model and the prior which justifies the use of the “least-squares” terminology for this type of problem. We separately discuss linear and non-linear problems in the next sections.

9 Linear problems

In the linear case, $\mathbf{g}(\mathbf{m}) = \mathbf{G}\mathbf{m}$ and the misfit function is quadratic in \mathbf{m} . As a result, the posterior pdf is Gaussian, and we can write

$$\sigma_M(\mathbf{m}) = k \exp\left(-\frac{1}{2}(\mathbf{m} - \tilde{\mathbf{m}})^H \tilde{\mathbf{C}}_M^{-1}(\mathbf{m} - \tilde{\mathbf{m}})\right)$$

where

$$\begin{aligned} \tilde{\mathbf{m}} &= (\mathbf{G}^H \mathbf{C}_D^{-1} \mathbf{G} + \mathbf{C}_M^{-1})^{-1} (\mathbf{G}^H \mathbf{C}_D^{-1} \mathbf{d} + \mathbf{C}_M^{-1} \mathbf{m}_0) \\ &= \mathbf{m}_0 + (\mathbf{G}^H \mathbf{C}_D^{-1} \mathbf{G} + \mathbf{C}_M^{-1})^{-1} \mathbf{G}^H \mathbf{C}_D^{-1} (\mathbf{d} - \mathbf{G}\mathbf{m}_0) \\ &= \mathbf{m}_0 + \mathbf{C}_M \mathbf{G}^H (\mathbf{G} \mathbf{C}_M \mathbf{G}^H + \mathbf{C}_D)^{-1} (\mathbf{d} - \mathbf{G}\mathbf{m}_0) \end{aligned}$$

and

$$\begin{aligned} \tilde{\mathbf{C}}_M &= (\mathbf{G}^H \mathbf{C}_D^{-1} \mathbf{G} + \mathbf{C}_M^{-1})^{-1} \\ &= \mathbf{C}_M - \mathbf{C}_M \mathbf{G}^H (\mathbf{G} \mathbf{C}_M \mathbf{G}^H + \mathbf{C}_D)^{-1} \mathbf{G} \mathbf{C}_M. \end{aligned}$$

To assist in remembering these expressions, consider that we can write

$$\tilde{\mathbf{C}}_M = (\mathbf{I} + \mathbf{G}^A \mathbf{G})^{-1} \mathbf{C}_M$$

so that

$$\begin{aligned} \tilde{\mathbf{m}} &= \mathbf{m}_0 + \tilde{\mathbf{C}}_M \mathbf{G}^H \mathbf{C}_D^{-1} (\mathbf{d} - \mathbf{G}\mathbf{m}_0) \\ &= \mathbf{m}_0 + (\mathbf{I} + \mathbf{G}^A \mathbf{G})^{-1} \mathbf{G}^A (\mathbf{d} - \mathbf{G}\mathbf{m}_0) \\ &= \mathbf{m}_0 + \mathbf{G}^{\tilde{A}} (\mathbf{d} - \mathbf{G}\mathbf{m}_0). \end{aligned}$$

where $\mathbf{G}^{\bar{A}}$ is an adjoint-like operator defined using the covariance $\tilde{\mathbf{C}}_M$, instead of the covariance \mathbf{C}_M . This equation states that the difference between the estimate and the prior is the difference between the data and the prior-predicted-data mapped through an adjoint-like operator of \mathbf{G} that uses $\tilde{\mathbf{C}}_M$ instead of \mathbf{C}_M . In terms of the actual adjoint of \mathbf{G}

$$\mathbf{G}^{\bar{A}} = (\mathbf{I} + \mathbf{G}^A \mathbf{G})^{-1} \mathbf{G}^A.$$

If there is some flexibility in the choice of \mathbf{G} , then it is better to choose \mathbf{G} so that $(\mathbf{I} + \mathbf{G}^A \mathbf{G})^{-1}$ has eigenvalues as small as possible (or so that its diagonal entries are as small as possible).