Optimal Interpolation and/or “kriging” consist in determining the best linear estimate in the “least square sense” for locations \( x_i \) where you have no measurements:

**Example 1:**

Collected SST

\[
Z_i = a Z_1 + b Z_2 + c Z_3
\]

Estimated SST

```matlab
X=1e3.*[40
-45
45];
Y=1e3.*[60
-30
-70];
Z=[24.6
26.4
25.2];
%
XI=1e3*10;
YI=1e3*15;

ZI=?
```

Linear Estimate
Optimal Interpolation

General Case: N measurements

\[ \hat{Z}_i = \sum_{k=1}^{N} w_k Z_k \]

- Estimated data at location \((x_i, y_i)\)
- Linear coefficients: we attribute a "weight" to each collected data
- Measured data at locations \((x_k, y_k), k=1, \ldots, N\)

How do we calculate the linear coefficients, \(w_k\)?

minimizing the estimated error variance!

True Value  Estimated value

\[ \hat{s}_e^2 = \text{var} \left( Z_i - \hat{Z}_i \right) \]

error
Optimal Interpolation

\[ \hat{s}_e^2 = \text{var} \left( Z_i - \hat{Z}_i \right) \]

\[ \hat{s}_e^2(i) = \left\langle \left( Z_i - \hat{Z}_i \right)^2 \right\rangle \]

where \( \left\langle \cdot \right\rangle \) is the expected mean drawn form a "large" number of realizations or some large statistical population.

NOTE: This “large” statistical population is usually not sampled, meaning “a priori information” must be given!

\[ \hat{s}_e^2(i) = \left\langle \left( Z_i - \sum_{k=1}^{N} w_k Z_k \right)^2 \right\rangle \]

\[ \hat{s}_e^2(i) = \left\langle Z_i Z_i \right\rangle - 2 \sum_{k=1}^{N} w_k \left\langle Z_i Z_k \right\rangle + \sum_{k=1}^{N} \sum_{m=1}^{N} w_k w_m \left\langle Z_k Z_m \right\rangle \]
Optimal Interpolation

\[ \hat{s}_e^2(i) = \langle Z_i Z_i \rangle - 2 \sum_{k=1}^{N} w_k \langle Z_i Z_k \rangle + \sum_{k=1}^{N} \sum_{m=1}^{N} w_k w_m \langle Z_k Z_m \rangle \]

We solve for the unknowns \( w_k \) minimizing the error variance (similar to the least square problem)

\[ \frac{\partial \hat{s}_e^2(i)}{\partial w_k} = 0, \quad k = 1, 2, \ldots, N \]

\[ \sum_{m=1}^{N} w_m \langle Z_k Z_m \rangle = \langle Z_i Z_k \rangle \quad k = 1, 2, \ldots, N \]

N unknowns, N linear equations! 

Do we still have a problem?
Optimal Interpolation

Problems:
- The true values $Z_i$ is not known
- We have usually not sampled a large statistical population to estimate the $<Z_iZ_k>$.

We add a priori information to the problem:
- a covariance model, from which we are getting the Covariance Matrix

$$\langle Z'_k Z'_m \rangle = C(d_{km})$$

The covariance between data collected at two locations depends on the distance that separates the two location

NOTE: $Z'_k$ is the anomaly (a spatial trend has been removed)

$$Z'_k = Z_k - \overline{Z}_k$$

$\overline{Z}'_k = 0$

\(\overline{\text{...}}\) refers to a spatial trend or mean
We go back to our linear system of $N$ equations:

$$\sum_{m=1}^{N} w_m \langle Z_k \bar{Z}_m \rangle = \langle Z_i \bar{Z}_k \rangle \quad k = 1, 2, \ldots, N$$

We rewrite the problem using the anomalies (removing the mean or trend):

$$\sum_{m=1}^{N} w_m \langle \bar{Z}_k \bar{Z}_m \rangle = \langle \bar{Z}_i \bar{Z}_k \rangle \quad k = 1, 2, \ldots, N$$

We define: $C(d_{km}) \equiv C_{km}$

$$\sum_{m=1}^{N} w_m C_{km} = C_{ki} \quad k = 1, 2, \ldots, N$$
\[
\sum_{m=1}^{N} w_m C_{km} = C_{ki} \quad k = 1, 2, \ldots, N
\]

In Matrix Form:
\[
C_{km} \mathbf{w} = C_{ki}
\]

Estimated-data covariance vector

Data covariance matrix (a priori information)
N x N if data at N locations

Linear coefficients ("weights") for the optimal interpolation at location i
\[ C_{km} \mathbf{w} = C_{ki} \]

\[ \mathbf{w} = C_{km}^{-1} C_{ki} \]

\[ \hat{Z}_i = \sum_{k=1}^{N} w_k Z_k = \mathbf{w}^T \mathbf{Z}_k \]

\[ \hat{Z}_i = \left( C_{km}^{-1} C_{ki} \right)^T \mathbf{Z}_k \]

**N\times N** data covariance matrix

**N\times 1** estimated-data covariance vector

**1\times N** "line" vector of weights

**N\times 1** vector of data

Estimated value at one location \((x_i, y_i)\)
Computing the error in the mean square sense:

\[ \hat{s}_{e(i)}^2 = \langle Z_i Z_i \rangle - 2 \sum_{k=1}^{N} w_k \langle Z_i Z_k \rangle + \sum_{k=1}^{N} \sum_{m=1}^{N} w_k w_m \langle Z_k Z_m \rangle \]

Minimizing the mean square error has led to:

\[ \sum_{m=1}^{N} w_m \langle Z_k Z_m \rangle = \langle Z_i Z_k \rangle \quad k = 1, 2, \ldots, N \]

Therefore:

\[ \hat{s}_{e(i)}^2 = \langle Z_i Z_i \rangle - \sum_{k=1}^{N} w_k \langle Z_i Z_k \rangle \]

\[ \hat{s}_{e(i)}^2 = C(0) - \left( C_{km}^{-1} C_{ki} \right)^T C_{ki} \]
Note:

\[
\langle \hat{Z}_i^2 \rangle = \left\langle \left( \sum_{k=1}^{N} w_k Z_k \right)^2 \right\rangle = \sum_{k=1}^{N} \sum_{m=1}^{N} w_k w_m \langle Z_k Z_m \rangle
\]

\[
\langle \hat{Z}_i^2 \rangle = \sum_{k=1}^{N} w_k \langle Z_i Z_k \rangle
\]

Therefore:

\[
\hat{S}_e^2(i) = \langle Z_i Z_i \rangle - \langle \hat{Z}_i^2 \rangle
\]

1) If no data available:

\[
\hat{Z}_i = \langle Z_i \rangle = Z_0
\]

Best estimate is the expected value

\[
\hat{s}^2(i) = C(0)
\]

Expected error is the Covariance of the data at zero lag
2) If data available:

\[ \hat{s}_e^2(i) = \langle Z_i Z_i \rangle - \left\langle \hat{Z}_i^2 \right\rangle \]

The presence of data results in a reduction of the error by:

\[ \sum_{k=1}^{N} w_k C_{ik} > 0 \]

\[ \hat{s}_e^2(i) = C(0) - \left( C^{-1}_{km} C_{ki} \right)^T C_{ki} \]

- Error variance estimate (or square error or "confidence interval")
- Covariance of the data at zero lag expected error if no observations
- Weights: nearby data that is expected to covary positively are assigned larger weights
We relax the assumption that we know the mean drift:

$$\bar{Z}_i(x_i) = Z_0$$

To do so, we must add a constraint. In order to keep the estimate unbiased, this constraint must be:

$$\sum_{k=1}^{N} w_k = 1$$

The problem now results in minimizing a function of $N$ variables, the error variance that is function of the $w_k$, with a constraint on the $w_k$. Mathematically, we look for $w_k$ that satisfy:

$$\frac{\partial}{\partial w_k} \left( \hat{s}_e^2(i) - \lambda \left( \sum_{k=1}^{N} w_k - 1 \right) \right) = 0$$
\[ \sum_{m=1}^{N} w_m \langle Z'_k Z'_m \rangle + \lambda = \langle Z'_i Z'_k \rangle \quad k = 1, 2, \ldots, N \]

In Matrix form:

\[
\begin{bmatrix}
C_{1i} \\
C_{2i} \\
\vdots \\
C_{ki} \\
C_{Ni} \\
1
\end{bmatrix}
= 
\begin{bmatrix}
C_{11} & C_{12} & \cdots & C_{1k} & \cdots & C_{1N} & 1 \\
C_{21} & C_{22} & \cdots & C_{2k} & \cdots & C_{2N} & 1 \\
\vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots \\
C_{ki} & C_{k2} & \cdots & C_{kk} & \cdots & C_{kN} & 1 \\
\vdots & \vdots & \cdots & \vdots & \ddots & \vdots & \vdots \\
C_{Ni} & C_{N2} & \cdots & C_{Nk} & \cdots & C_{NN} & 1 \\
1 & 1 & \cdots & 1 & \cdots & 1 & 0
\end{bmatrix}
\begin{bmatrix}
w_1 \\
w_2 \\
\vdots \\
w_k \\
w_N \\
\lambda
\end{bmatrix}
\]
Example:

```matlab
%% Collected Data
X = 1e3 .* [40 -45 45];
Y = 1e3 .* [60 -30 -70];
Z = [24.6 26.4 25.2];

%% Location where we want to estimate the data
XI = 1e3 * 10;
YI = 1e3 * 15;
```
% Covariance data matrix
for k=1:3
   for m=1:3
      dkm(k,m)=sqrt(abs(X(k)-X(m))^2+abs(Y(k)-Y(m))^2);
   end
end

Ckm=(1-0.01)*dkm;

for k=1:3
   dki(k,1)=sqrt(abs(X(k)-XI)^2+abs(Y(k)-YI)^2);
end

Cki=(1-0.01)*dki;

% w=inv(Ckm)*Cki

% BE CAREFUL: BAD ESTIMATION if means is not substracted
ZI=w'*Z;

% GOOD ESTIMATION if the mean is removed
Zprim=Z-mean(Z);
ZIprim=w'*Zprim;
ZIgood=mean(Z)+ZIprim;

ZIgood = 

   25.3011
   ZI = 

   22.9004
%% Ordinary krigging
Ckmok=zeros(4,4);
Ckmok(1:3,1:3)=Ckm;
Ckmok(4,1:3)=1;
Ckmok(1:3,4)=1;

Ckiok=ones(4,1);
Ckiok(1:3,1)=Cki;

wok=inv(Ckmok)*Ckiok;
ZIok=mean(Z)+wok(1:3)'*(Z-mean(Z));

ZIok =

  25.2944

Check that:

>> sum(wok(1:3))

ans =

    1.0000
IF the MEAN is not removed:

\[
\langle Z_k, Z_m \rangle = \langle Z_0^2 \rangle + \langle Z'_k, Z'_m \rangle
\]

\[
R_{km} \quad C_{km}
\]

%% Papoulis avec moyenne différent de zéros
Rkm=ones(3,3);
Rkm(1:3,1:3)=Ckm+mean(Z)*mean(Z);
Rki(1:3,1)=Cki+mean(Z)*mean(Z);
wpap=inv(Rkm)*Rki;
ZIpap=wpap'*Z

ZIpap =

22.9201

BAD ESTIMATE

There is a better way of estimating the data, adding a variable \( Z_{N+1} = 1 \) (Papoulis, p393)
Rkmpap = ones(4, 4);
Rkmpap(1:3, 1:3) = Ckm + mean(Z) * mean(Z);
Rkmpap(4, 1:3) = mean(Z);
Rkmpap(1:3, 4) = mean(Z);

Rkipap = ones(4, 1);
Rkipap(1:3, 1) = Cki + mean(Z) * mean(Z);
Rkipap(4, 1) = mean(Z);

wpap = inv(Rkmpap) * Rkipap;
Zpap = ones(4, 1);
Zpap(1:3) = Z;
Zpap(4) = 1;
ZIpap2 = wpap' * Zpap

ZIpap2

25.3011
Example 2: AVHRR SST image with NaN values (due to clouds)

114 x 114 SST image with NaN values

Objective: estimate the SST at locations where we have no data
1. Load initial data
   here, 3 SST images...
clear all
avhrr_init;

Get rid of non physical values
X1sst(X1sst>30)=NaN;
X2sst(X2sst>30)=NaN;
X3sst(X3sst>30)=NaN;

Define or load the grid
lond=-69:(10/113):-59;
latd=27:(10/113):37;

Extract Subdomain
lonmin=-69;
lonmax=-63;
latmin=30;
latmax=33;
imin=find(abs(lond-lonmin)==min(abs(lond-lonmin)));
imax=find(abs(lond-lonmax)==min(abs(lond-lonmax)));
jmin=find(abs(latd-latmin)==min(abs(latd-latmin)));
jmax=find(abs(latd-latmax)==min(abs(latd-latmax)));

londsub=lond(imin:imax);
latdsub=latd(jmin:jmax);
Z=X1sst(imin:imax,jmin:jmax);
londsub=lond;
latdsub=latd;
Z=X3sst;

[Londsub,Latdsub]=ndgrid(londsub,latdsub);

% Some plots
% Note the difference between pcolor and imagesc
figure;
load purp_map
colormap(purp_map);
pcolor(Londsub,Latdsub,Z);shading flat;
colorbar;
2. Separate the GOOD data (used for the optimal interpolation) from the BAD data.
   Our objective is to use a linear combination of the GOOD data and a data value at the locations where we have BAD data.

```matlab
igood=find(~isnan(Z));
ibad=find(isnan(Z));

3. Estimate the data-data "distance" covariance matrix

```matlab
for i=1:length(igood)
    for j=1:length(igood)
        dkm(i,j)= spheric_dist(Latdsub(igood(i)),Latdsub(igood(j)),... 
                               Londsub(igood(i)),Londsub(igood(j)));
    end
end

```matlab

```matlab
% We choose a covariance model:
% C(d)= exp(-d^2/d0^2)

d0=20e3;
Ckm=exp(-dkm.^2/d0^2)+eye(size(dkm));
Zoi=Z;

Ckmok=zeros(size(Ckm)+1);
Ckmok(1:size(Ckm,1),1:size(Ckm,2))=Ckm;
Ckmok(size(Ckm,1)+1,1:size(Ckm,2))=1;
Ckmok(1:size(Ckm,1),size(Ckm,2)+1)=1;

% Covariance between your data points and the estimated data
Zek=Z;
```
\% 4. Data -estimated covariance matrix
\%
dkib=spheric_dist_oil(Latdsub(igood),Latdsub(iband),
    
    Londsub(igood),Londsub(iband));
Ckib=exp(-dkib.^2/d0^2);

Ckibok=ones(size(Ckib,1)+1,size(Ckib,2));
Ckibok(1:size(Ckib,1),:)=Ckib;
\%
5. Computation of the linear coefficients
\% There are Ngood coefficients for each of the Nbad values
wb=inv(Ckm)*Ckib;
w=Ckm\backslash Cki;

\% 6. Estimation of the data where it was missssing
Zoi(iband)=wb'*(Z(igood)-mean(Z(igood)))+mean(Z(igood));

    Zok(iband)=wbok(1:length(igood),:)'*... 
    (Z(igood)-mean(Z(igood)))+mean(Z(igood));

\% 7. Estimation of the error
C0=1;
Se(iband)=C0-diag(wb'*Ckib);
Se(igood)=0;
Se=reshape(Se,[size(Z,1) size(Z,2)]);
The full image after filling the gaps with the ordinary krigging method.
The error is maximum at the points where we estimated data. The error is maximum at points located far away (beyond d0) from good data, therefore it depends on d0. The amplitude of the error also depends on your variance (C0).
To do:

Vary the parameter d0 chosen in the covariance model: choose large d0 and small d0

How can we “objectively” choose a covariance model and the parameters associated to it?
Variance and Covariance

Variance of ONE variable $Z$:

$$s^2 = \frac{1}{(N - 1)} \sum_{k=1}^{k=N} (Z_i - \bar{Z})^2$$

It is mean of square deviation of your data

Covariance of TWO variables $Z^j$ and $Z^k$:

$$s^2_{jk} = \frac{1}{(N - 1)} \sum_{k=1}^{k=N} (Z^j_i - \bar{Z}^j)(Z^k_i - \bar{Z}^k)$$
Correlation

Correlation coefficient of TWO variables \(Z^j\) and \(Z^k\):

\[
 r_{jk} = \frac{s_{jk}^2}{s_j s_k}
\]

It is a statistical measure of whether two variables are linearly related to each other

\[-1 < r_{jk} < 1\]

\[
 r_{jk} = 1 \quad \text{perfectly positively correlated data}
\]

\[
 r_{jk} = 0 \quad \text{uncorrelated data}
\]

\[
 r_{jk} = -1 \quad \text{perfectly negatively correlated data}
\]

Of course, if \(j=k\), the two variables are identical, and they are perfectly correlated to each other
Auto Covariance

Time Series

You want to know to what extent the measurement you collected at time \( t+\tau \) depends on the measurement you collected at time \( t \).

Consider you have a times series of length \( N \), with equally “time” spaced measurement \( Z_k \). We define the “autocovariance” (because the two variables are the same) at lag \( n\Delta T \):

\[
C(n) = \frac{1}{(N-n)} \sum_{k=1+n}^{k=N} (Z_k - \bar{Z}_{1+n}^N) (Z_{k-n} - \bar{Z}_1^{N-n})
\]

\[
\bar{Z}_{1+n}^N = \frac{1}{(N-n)} \sum_{m=1+n}^{m=N} Z_m
\]

\[
\bar{Z}_1^{N-n} = \frac{1}{(N-n)} \sum_{m=1}^{m=N-n} Z_m
\]

Mean of the \((N-n)\) data values starting from the \( n+1 \) measurement.

Mean of the \((N-n)\) data values starting from the 1st measurement.
Autocorrelation at zero lag:

\[
C(n = 0) = \frac{1}{(N)} \sum_{k=1}^{k=N} \left( Z_k - \bar{Z} \right) \left( Z_k - \bar{Z} \right)
\]

\[
= \frac{1}{(N)} \sum_{k=1}^{k=N} Z_k^2 - 2\bar{Z} \frac{1}{(N)} \sum_{k=1}^{k=N} Z_k + \bar{Z}^2
\]

\[
= \frac{1}{(N)} \sum_{k=1}^{k=N} Z_k^2 - \bar{Z}^2
\]

only difference with variance is the division by instead of \((N-1)!\)
\[ C(0) \]

\[ C(1) \]

\[ C(n) \]

\%AutoCovariances

cov0 = mean(Z.^2) - mean(Z).^2;
cov(1) = mean(Z(2:N) .* Z(1:N-1)) - mean(Z(2:N)) * mean(Z(1:N-1));
cov(n) = mean(Z(n+1:N) .* Z(1:N-n)) - mean(Z(n+1:N)) * mean(Z(1:N-n));
Czz = xcov(Z, 'unbiased');

2N+1 values with the zero lag at the middle:
Czz(N) = cov0
Notes:

**Covariance** has units!

You may wish to compare several time series with different units. In that case, we use **correlation**.

Correlation is a normalized form of covariance.

**Autocorrelation** (when the two variables are the same):

\[
R(n) = \frac{C(n)}{s^2}
\]

At zero lag, the autocorrelation is 1.

In case of non periodic signal, it rapidly decreases as time increases.

\(C(n)\) for \(n > N/5\) usually not trustable.
Example 1: Periodic signal

```matlab
%% Covariance
% Exemple 1

t=0:3600:14*3600*24;
T1=12*3600;
eta=sin(2*pi/T1*t);
figure;
plot(t./24/3600,eta);
set(gca,'xtick',[1:14]);
print -depsc2 covariance-exemple1.eps
```

How does the covariance function look like?
Matlab solution:

```matlab
[cova,tau]=xcov(eta,'unbiased');
figure;
plot(tau(length(t):end),cov(length(t):end));
print -depsc2 covariance-exemple1-1.eps
```

Analytic solution:

\[
C(\tau) = \frac{1}{T} \int_0^T \sin \left( \frac{2\pi}{T} t \right) \sin \left( \frac{2\pi}{T} (t - \tau) \right) \, dt
\]

\[
= \frac{1}{2} \cos \left( \frac{2\pi}{T} \tau \right)
\]
X axis is the lag \( n \)
To convert it into unit: \( n \Delta t \)
In our example, \( \Delta t = 1 \) hour

The covariance of a periodic signal is periodic!
Example 2: Mimicking a tidal signal ...

How does the covariance function look like?

```matlab
%%
t=0:3600:14*3600*24;
T1=12*3600;
T2=14*24*3600
eta=sin(2*pi/T1*t)+sin(+2*pi/T2*t);
figure;
plot(t./24/3600,eta);
set(gca,'xtick',[1:14]);
print -depsc2 mimick-of-tidal.eps
```
[cov, tau] = xcov(eta, 'unbiased');
figure;
plot(tau(length(t):end), cov(length(t):end));
print -depsc2 mimick-of-tidal-cov.eps
You want to know to what extent the data you collected at two locations separated by a distance $d$ depend on each other.

Consider data collected at $N$ locations, equally spaced. We define the “distance autocovariance” (because the two variables are the same) at lag $nd$:

$$C(d) = \frac{1}{N(d)} \sum_{k=1}^{N(d)} \left[ \left( Z(X_k) - \overline{Z}(X) \right) \times \ldots \right] \left( Z(X_k + du) - \overline{Z}(X) \right) .$$

$C(d)$, computed from experimental data and plotted as a function of the distance lag $d$, is named the “variogram”.
“variograms” (covariance) models commonly used for optimal interpolation

Linear:
\[ C(d) = C_0 + b \times d \]

Exponential:
\[ C(d) = C_0 + (C_\infty - C_0) \left( 1 - \exp\left( -\frac{d}{d_0} \right) \right) \]

Gaussian:
\[ C(d) = C_0 + (C_\infty - C_0) \left( 1 - \exp\left( -\frac{d^2}{d_0^2} \right) \right) \]

Spherical:
\[ C(d) = C_0 + (C_\infty - C_0) \left[ \frac{3 \, d}{2 \, d_0} - \frac{1 \, d^3}{2 \, d_0^3} \right] \quad \text{for} \quad d \leq d_0 \]
\[ C(d) = C_\infty \quad \text{for} \quad d > d_0 \]
• C0: Unresolved variance, or measurement error found as the intercept of the variogram at \(d=0\). This is also named the “nugget”

• Cinfty: variance between uncorrelated data pairs as lag \((d)\) goes to infinity; This is also named the “sill”

• d0: distance beyond which data points are no longer correlated. This is also named the “range”
Problem:

Choose the SST image chosen as an example for the optimal interpolation

Compute the “lag autocovariance” function, also named “variogram” of this image

Fit one of the covariance model to your “empirical” variogram

Apply it to your optimal interpolation.
```matlab
%%
d=0:20e3:500e3;
for i=1:length(d)
    clear idkm idkmx idkmy
    if (d(i)~=0)
        idkm=find(dkm<d(i)+dx & dkm>d(i)-dx);
        [idkmx,idkmy]=ind2sub(size(dkm),idkm);
        covc(i)=sum((Zgood(idkmx)-mean(Zgood)).*(Zgood(idkmy)-mean(Zgood)))./length(idkmx);
    else
        idkm=find(dkm==0);
        [idkmx,idkmy]=ind2sub(size(dkm),idkm);
        covc(i)=sum((Zgood(idkmx)-mean(Zgood)).*(Zgood(idkmy)-mean(Zgood)))./length(idkmx);
    end
end
figure;plot(covc);
```

The diagram shows the covariance function $Cov(d) = \langle Z_i Z_j \rangle$ with $\|\mathbf{r}_i - \mathbf{r}_j\| = d$. The graph presents $Cov(d)(\cdot K)^2$ on the y-axis against $d$ in kilometers (km) on the x-axis, illustrating how the covariance decreases with increasing distance $d$. The curve demonstrates the typical exponential decay of spatial correlation in spatial statistics and geostatistics.
\[ Cov(d) = \langle Z_i Z_j \rangle \text{ with } \|r_i - r_j\| = d \]

\[ \text{>> disp(p(1))} \]
\[ \quad 0.6259 \]

\[ \text{>> disp(p(2))} \]
\[ \quad 0.0770 \]

\[ \text{>> disp(p(3))} \]
\[ \quad 1.4709 \times 10^5 \]
TIPS:

Check the “physical” aspect of your estimated data
   Is the estimated data in a “respectable” range of values ?
   If not, what can I do ?
      Modify the parameters of the covariance model ...
      Check the covariance matrix (see on of the next point)

Check the sensitivity of your method to the covariance model.

Check the sensitivity of your method to the parameters of the covariance model (nugget, sill, range)

Remember that the results of the method depends on the invertibility of the covariance matrix. This means that the matrix must not be singular.
Why should it be singular at the first place? The data that you use to build the “variogram” may have some errors. Taking into account this error can help making the covariance matrix less singular

Check the error given by the method.
NOTES:

Measured data points will influence the estimated data within the range d0. Beyond d0, they will not estimate the solution of the interpolation.

If you have no data, or estimated data located beyond d0, then the best estimate given by the method will be the mean, Z0. And the error will be maximal!
Adding some errors to the measurements

We consider a sequence of two random variable, \( Z_k \) (collected data) and \( Z^s_k \) (signal) related by the following relation:

\[
Z_k = Z^s_k + \nu_k
\]

Our purpose is the same as before: it consists in estimating some data \( Z_i \) using a linear combination of the \( Z_k \), that minimizes the error in the least square sense.

\[
\hat{Z}_i = \sum_{k=1}^{N} w_k Z_k
\]

\[
\hat{s}_e^2 = var \left( Z_i - \hat{Z}_i \right)
\]
Adding some errors to the measurements

We consider a sequence of two random variables, \( Z_k \) (collected data) and \( Z^s_k \) (signal) related by the following relation:

\[
Z_k = Z^s_k + \nu_k
\]

collected data \hspace{1cm} “true” \hspace{1cm} noise

signal

Our purpose is the same as before: it consists in estimating some data \( Z_i \) using a linear combination of the \( Z_k \) that minimizes the error in the least square sense.

\[
\hat{Z}_i = \sum_{k=1}^{N} w_k Z_k
\]

\[
\hat{s}_c^2 = \text{var} \left( Z_i - \hat{Z}_i \right)
\]

function to minimize
Adding some errors to the measurements

\[ \hat{s}_e^2(i) = \left\langle \left( Z_i - \sum_{k=1}^{N} w_k Z_k \right)^2 \right\rangle \]

\[ \hat{s}_e^2(i) = \left\langle \left( Z_i - \sum_{k=1}^{N} w_k (Z_k^s + \nu_k) \right)^2 \right\rangle \]

\[ \hat{s}_e^2(i) = \langle Z_i Z_i \rangle - \cdots - 2 \sum_{k=1}^{N} w_k \langle Z_i (Z_k^s + \nu_k) \rangle + \cdots \]

\[ \frac{2 \sum_{k=1}^{N} w_k \sum_{m=1}^{N} w_m \langle (Z_k^s + \nu_k)(Z_m^s + \nu_m) \rangle}{\sum_{k=1}^{N} \sum_{m=1}^{N} w_k w_m} \]
Adding some errors to the measurements

Minimizing the error, looking for a local extremum:

\[
\frac{\partial \hat{s}_e^2(i)}{\partial w_k} = 0, \quad k = 1, 2, \ldots, N
\]

\[
\sum_{m=1}^{N} w_m \left\langle \left( Z^s_k + \nu_k \right) \left( Z^s_m + \nu_m \right) \right\rangle = \left\langle Z_i \left( Z^s_k + \nu_k \right) \right\rangle \quad k = 1, 2, \ldots, N
\]

We can simplify the problem if we make the assumption that:

1) the noise attached to the measurement collected at location \( r_m \) is not correlated to the noise attached to the data collected at location \( r_k \)

\[
\left\langle \nu_k \nu_m \right\rangle = \delta_{k,m} \nu^2
\]

kronecker symbol

noise variance

2) the signal is uncorrelated to the noise:

\[
\left\langle Z_k \nu_m \right\rangle = 0
\]
Adding some errors to the measurements

\[
\sum_{m=1}^{N} w_m \left( \langle Z_k^S Z_m^S \rangle + \langle \nu_k \nu_m \rangle \right) = \langle Z_i Z_k^S \rangle \quad k = 1, 2, \cdots , N
\]

\[
\sum_{m=1}^{N} w_m C_{km} + \delta_{km} \nu^2 = C_{ki} \quad k = 1, 2, \cdots , N
\]

\[
\begin{bmatrix}
C_{1i} \\
C_{2i} \\
\vdots \\
C_{ki} \\
\vdots \\
C_{Ni}
\end{bmatrix} = \begin{bmatrix}
C_{11} + \nu^2 & C_{12} & \cdots & C_{1k} & \cdots & C_{1N} \\
C_{21} & C_{22} + \nu^2 & \cdots & C_{2k} & \cdots & C_{2N} \\
\vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\
C_{k1} & C_{k2} & \cdots & C_{kk} + \nu^2 & \cdots & C_{kN} \\
\vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\
C_{N1} & C_{N2} & \cdots & C_{Nk} & \cdots & C_{NN} + \nu^2
\end{bmatrix} \times \begin{bmatrix}
w_1 \\
w_2 \\
\vdots \\
w_k \\
\vdots \\
w_N
\end{bmatrix}
\]
Adding some errors to the measurements

\[ \sum_{m=1}^{N} w_m \left( \langle Z_k^s Z_m^s \rangle + \langle \nu_k \nu_m \rangle \right) = \langle Z_i Z_k^s \rangle \quad k = 1, 2, \ldots, N \]

\[ \sum_{m=1}^{N} w_m C_{km} + \delta_{km} \nu^2 = C_{ki} \quad k = 1, 2, \ldots, N \]

\[
\begin{bmatrix}
C_{1i} \\
C_{2i} \\
\vdots \\
C_{ki} \\
C_{Ni}
\end{bmatrix}
= 
\begin{bmatrix}
C_{11} + \nu^2 & C_{12} & \ldots & C_{1k} & \ldots & C_{1N} \\
C_{21} & C_{22} + \nu^2 & \ldots & C_{2k} & \ldots & C_{2N} \\
\vdots & \vdots & \ddots & \vdots & \ldots & \vdots \\
C_{k1} & C_{k2} & \ldots & C_{kk} + \nu^2 & \ldots & C_{kN} \\
\vdots & \vdots & \ldots & \vdots & \ddots & \vdots \\
C_{N1} & C_{N2} & \ldots & C_{Nk} & \ldots & C_{NN} + \nu^2
\end{bmatrix}
\begin{bmatrix}
w_1 \\
w_2 \\
\vdots \\
w_k \\
w_N
\end{bmatrix}
\]

\[ w_i = (C_{km} + \delta_{km} \nu^2)^{-1} C_{ki} \]
Adding some errors to the measurements

% We choose a covariance model:
% \( C(d) = \exp\left(-d^2/d0^2\right) \)
% \( d0 = 20e3 \);
C0b = 0.68;
Cinfb = 0.07;
d0b = 300e3;
nu2 = 0.24;

% \( C_{km} = \exp\left(-d_{km}^2/d0^2\right) + \text{eye(size(dkm))} \);
Ckm = expmod(dkm, [C0b, Cinfb, d0b]);
Ckm = Ckm + nu2 * eye(size(dkm));
Adding some errors to the measurements

\[
\hat{s}_e^2(i) = \langle Z_i Z_i \rangle - 2 \sum_{k=1}^{N} w_k \langle Z_i Z_k \rangle + \sum_{k=1}^{N} \sum_{m=1}^{N} w_k w_m \langle Z_k Z_m \rangle
\]

Minimizing the mean square error has led to:

\[
\sum_{m=1}^{N} w_m \langle Z_k Z_m \rangle = \langle Z_i Z_k \rangle \quad k = 1, 2, \ldots, N
\]

Therefore:

\[
\hat{s}_e^2(i) = \langle Z_i Z_i \rangle - \sum_{k=1}^{N} w_k \langle Z_i Z_k \rangle
\]

\[
\hat{s}_e^2(i) = C_{ii} + \nu^2 \delta_{ii} - w_i^T C_{ki} \quad \text{avec} \quad w_i = \left( C_{km} + \delta_{km} \nu^2 \right)^{-1} C_{ki}
\]
SST (°C)

Error (mean square sense) in (°C)²
To be done next year:

Show that the error is orthogonal to the data, so that extremum that is found corresponds to a minimum

Take an example from real data (fromvar, or TP rade)