

**Dynamic Data Assimilation**  
**Prof. S Lakshmivarahan**  
**School of Computer Science**  
**Indian Institute of Technology, Madras**

**Lecture - 30**  
**Optimal interpolations**

(Refer Slide Time: 00:21)

### A HISTORICAL VIEW

- During 1960s and 1970s, operational centers in the USA, Sweden, Japan and others routinely used the iterative schemes
- Also known as successive error correction
- In the Soviet union (1960s) a technique called optimal interpolation (OI) was championed by Lev Gandin
- OI was independently developed by N. Wiener (1949) in the USA and Kolmogorov (1941) in the Soviet Union

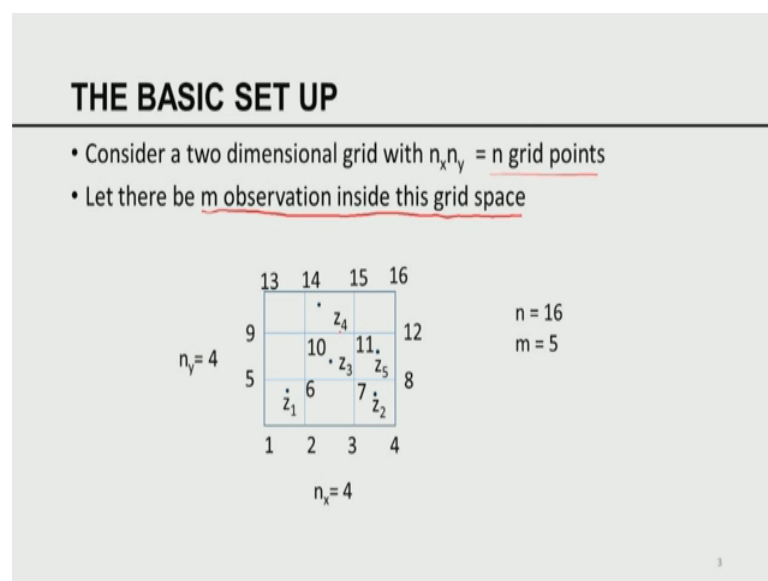
2

In this talk, we are going to be looking at a method that has come to be known as optimal interpolation. I would like to provide a quick historical perspective of this approach to prediction estimation based on a method that has come to be called optimal interpolation. During the 1960s and 70s operational centers in USA, Sweden, Japan and others routinely used iterative scheme like Batherson, Drues or Crosman type that we talked about in the last lecture. These methods have also come to be known as successive error correction, you can see  $z - h^T x_k$  in the iterative scheme, which we can think of in innovation is also can be thought of as the error. And by iteratively performing the update we are trying to transfer the information from the observation network to the computational network, so that method that class of methods crosman types schemes have also come to be called successive error correction method

Then this was going on in USA, Sweden, Japan rather countries, in Soviet Union in 1960s a technique called optimal interpolation was championed by Gandin - Lev Gandin. Optimal interpolation was developed earlier independently by Norbert wiener around

1940s in the USA and Kolmogorov in the Soviet Union the story goes that Norbert Wiener developed this method in the early 40s, but he did this work under a defense contract, so it was classified. And by the time he could publish it, it has to be unclassified. So, it took several years before the first classification of Wiener's ideas were known to the public. So, the publication date the open publication date is around 1949, but here also independently developed in the early 40s not knowing that Kolmogorov was also working on similar problems. This goes to show big minds think similarly even though they have been working in geographically distinct locations.

(Refer Slide Time: 03:14)



What is the basic idea of this method called optimal interpolation. I am going to illustrate this notion of optimal interpolation using a simple 2D grid problem. It can be extended to 3D grid as well. Consider a two-dimensional grid with  $n_x$  times  $n_y$  number of points;  $n_x$  is the number of points along the  $x$  direction;  $n_y$  is the number of points along the  $y$  direction,  $n$  is the total number of grid points. Let that be  $m$  observations inside this grid space. As an example I had  $n_x = 4$ ,  $n_y = 4$ . So,  $n$  is 16, I have a set of five observation  $z_1$ ,  $z_2$ ,  $z_3$ ,  $z_4$  and  $z_5$ ; we can think of this observation as scalar observation temperature, pressure, humidity or concentration of some chemical whatever quantity filtered by essentially it is a scalar.

(Refer Slide Time: 04:25)

## CASE – I PERFECT OBSERVATION

- Consider a field variable of interest – temperature pressure, etc
- Consider a time  $k$  and let
$$\mathbf{Z}_k = (Z_{k_1}, Z_{k_2}, \dots, Z_{k_m})^T \in \mathbb{R}^m$$
be the vector of observation from the  $m$  observation stations at a given time  $k$
- Since there is an inherent variability,  $\mathbf{Z}_k$  is a random vector
- It's assumed that the observations are error-free

4

So, I am now going to consider two cases in our illustration. First, let us assume the observations are perfect, observations of perfect mean there is no noise. Why do we do even though we know observation generally are associated noise, I think it is good to get a grip on the idea by assuming that observations are perfect. So, the field variable of interest as I said is a scalar field variable, it could be temperature, pressure etcetera. Consider time  $k$  noon, January 1st 2016 as an example. So,  $k$  is a specific instant time. There are  $m$  observational occasions. So, at that time  $k$ , there are  $m$  observations of the scalar field variable temperature pressure whatever, and I have a  $m$  vector the components of the vector are  $Z_{k_1}, Z_{k_2}, \dots, Z_{k_m}$  is a  $m$ -dimensional vector at time  $k$   $\mathbf{Z}_k$ . These are the observations that are available from the  $m$  observation stations.

Now, the observation that we are going concerned with has a natural variability in itself. For example, if you consider the temperature at the downtown Paris on December on January 1st of every year there is a natural variation. They no two successive years will have the exactly the same temperature in downtown Paris at noontime on January 1st for the 100 years. So, you can think of these ah scalar variable to be naturally varying, this natural variability is captured or is described as a random process. So, given this natural availability we are going to consider  $\mathbf{Z}_k$  as a random vector. So,  $\mathbf{Z}_k$  is random not because of the observation errors,  $\mathbf{Z}_k$  is not because of the inherent natural variability the resulting from changes that are inherent to climate variables. It is assumed the observations are error free that is what perfect observations relate to.

(Refer Slide Time: 07:15)

## SPATIAL COVARIANCE OF OBSERVATION

- Assume that  $Z_k$  are drawn from a stationary distribution
- Let  $\bar{Z} = E(Z)$  and  $\tilde{Z} = Z - \bar{Z}$
- $E[\tilde{Z}] = 0$ ,  $Cov(Z) = E(\tilde{Z}\tilde{Z}^T) = C \in R^{m \times m}$
- It is assumed that  $\bar{Z}$  and  $C$  are known.
- In practice, given a long time series,  $\bar{Z}$  and  $C$  are computed using the samples

$$C_{ii} = \frac{1}{N} \sum_{k=1}^N (\tilde{Z}_i(k))^2$$

$$C_{ij} = \frac{1}{N} \sum_{k=1}^N \tilde{Z}_i(k) \tilde{Z}_j(k)$$

5

So, the whole method of Wiener and Kolmogorov which has come to be call optimal interpolation rests on the assumption that the properties of the random vector  $Z_k$  is stationary.

(Refer Slide Time: 07:34)

## COMPUTATION OF SPATIAL COVARIANCE

- Let  $\{Z_k \mid 1 \leq k \leq T\}$  be the time series of field variables
- Then  $\bar{Z} = \frac{1}{T} \sum_{k=1}^T Z_k$ , sample average  $\rightarrow$  (1)
- Let  $\tilde{Z}_k = Z_k - \bar{Z}$ , anomaly  $\rightarrow$  (2)
- Then  $C = \frac{1}{T} \sum_{k=1}^T \tilde{Z}_k \tilde{Z}_k^T \rightarrow$  (3)
- Assume,  $C$  is SPD
- NOTE: This matrix  $C$  captures the natural variations in the observations. This is not related to the observation covariance matrix

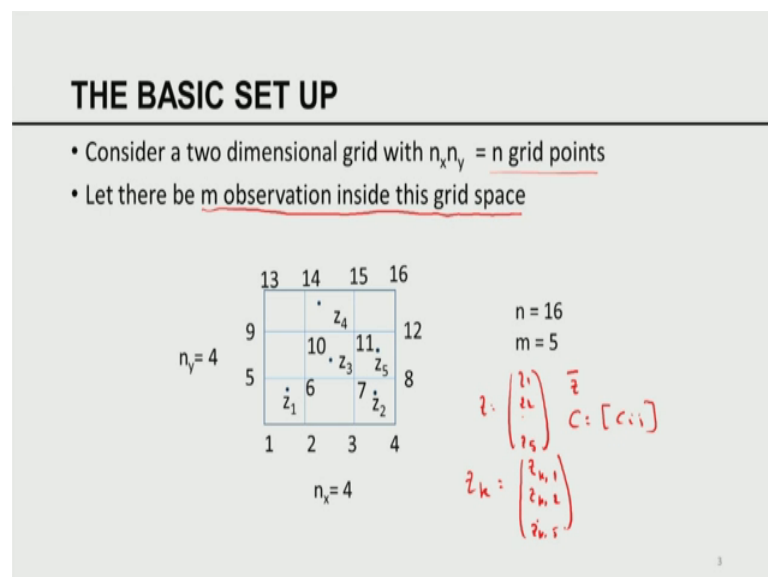
R
error

6

What does mean the random process that describes the temperature time series is I am sorry that is correct has a stationary distribution. So, what does it mean,  $Z$  is a random vector every random vector has an associated probability distribution. If that probability distribution is invariant in time, this is called a stationary distribution. If that probability

distribution changes in time is called a non-stationary distribution. A fundamental assumption is that the temperature in downtown Paris at noon on January 1st of every year is drawn from a stationary distribution. The stationary distribution is the one that describes the underlying natural variability. So, given this stationary distribution that is behind the realization of  $Z_k$ ;  $\bar{Z}$  is the mean of  $Z$ ;  $\tilde{Z}$  is the enamel  $Z$  minus  $\bar{Z}$ ; expected value of  $\bar{Z}$  is 0. The covariance of  $Z$  is given by expected value of  $\tilde{Z} \tilde{Z}^T$  and that I am going to call it  $C$ . So,  $C$  is a  $m$  by  $m$  matrix  $Z$  as  $C$  matrix talks about the covariance of the observations from different locations it is assumed that both the mean and the covariance of the observations are known.

(Refer Slide Time: 09:27)



So, let us go back. In this case, I have a grid with 16 points, I have 5 observations, I am now talking about observation vector. So, let us fix the time. So, I do not have to worry about  $k$ , let us fix the time  $k$  to be noon January 1st. So, at noon January 1st, I am getting five observations  $Z$  is equal to  $Z_1 Z_2$  up to  $Z_5$ . This will have a mean  $\bar{Z}$  this will have a covariance matrix  $C$ , the covariance matrix  $C$  will consist of  $C_{ij}$   $i$  runs from 1 to 5,  $j$  runs from 1 to 5. What is the meaning of  $C_{ij}$ ,  $C_{ij}$  is the correlation between the scalar variable at location  $i$  to the location  $j$ . So, it is assumed that this correlation is known that is a key to the methodology.

So, the whole question is we assume the distribution stationary, we may not be able to get a handle on the stationary distribution, but what is that one can do, I should be able to

estimate  $\bar{Z}$  and  $C$ . How do you estimate  $\bar{Z}$  and the covariance  $C$ . So, let us look at the station. So, let us take the time series of observations, so  $Z_1$ . So, if I now consider the time if I bring in the time now  $Z$  becomes  $Z_k$   $Z_k$  is equal to  $Z_{k1}$ ,  $Z_{k2}$  and  $Z_{k5}$ , therefore, if I consider station one I am going to have a time series of that measurement over a long period of time.

So, for example, I am trying to measure the temperature in downtown Paris at noon every day, so that time series is known. So, once that time series is known. So, what is that we are going to assume I am going to fix the location downtown Paris, I am going to fix a particular instance time namely noon of a given day. So, every day in downtown Paris, I am going to measure the temperature, and there is a record of it and that time series is made available to me that time series it is a long time series. Likewise so downtown Paris is only one observation station and I picked for an example in this illustration there are five such spatial observation towers or observation locations. From each location, at each instant of the day, at each hour of the day, there is going to be measurements of the scalar variable of interest. So, these sensors are going to spit out the values of the observed quantity and I am assuming I have already recorded, so that is the fundamental assumption.

So, given a long time series of observation from each of these locations, I can now compute the mean. So, I can say downtown Paris, what is the mean, so what is that I need to consider. What is the mean temperature in downtown Paris on noon January 1st what is the mean temperature in downtown Paris noon January second. So, day-by-day by processing the data, I should be able to compute the mean. Once I compute the mean, if I have time series, I should be able to compute an anomaly. Once they compute an anomaly, I can compute the variance of the measurements at a given location, this I can do for every location. So, I will have  $m$  locations, for a given time I will have  $m$  means, I will have  $m$  anomalies, I will have series of anomalies. Using the time series of anomalies, I can also now compute the covariance.

So, how do I compute the covariance? Let us say  $C_{ij}$ ,  $C_{ij}$  is equal to let us consider I am I am fixing the time right now. So, time case  $k$  is fixed. So, I have an anomaly at station  $i$  I have a time series of that with respect to  $k$ . So,  $Z_i$  tilde the  $k$  is the anomaly at the  $k$  instant in time at the  $i$ -th station  $Z_i$  tilde these are scalars I do not have to have even written like this, I simply can square the anomaly once again I can square the

anomaly. I can square the anomaly, I can sum this anomaly over  $k$  is equal to 1 to  $n$ , I can compute by this. And this will give me, this is not  $i j$  this is  $i i$ , so that is the variance at a given station. Now I would like to be able to compute the elements  $C_{ij}$  this is essentially equal to  $\frac{1}{n} \sum_{k=1}^n (Z_i - \bar{Z}_i)(Z_j - \bar{Z}_j)$  the product of the anomalies at station  $i$  and  $j$ , I am going to have to sum this up over  $k$  is equal to 1 to  $n$  over  $n$ . So, this is an estimate of the covariance between station  $i$  and station  $j$ , this is the variance of the station  $i$  these two together decides the elements of the covariance matrix.

So, what is the fundamental assumption, there are two key assumptions there are here. I am assuming the scalar field variable of interest such as pressure temperature whatever it is has a natural variability. This natural variability can be captured by a stationary distribution. Why stationarity in distribution analysis of non-stationary stochastic processes are extremely difficult there are very few results in the analysis for the analysis and quantification of properties of non-stationary stochastic processes. The only thing we know is to be able to pin down analyze characterize the properties of stationary stochastic processes.

So, in here  $Z_k$  is the observation vector at time  $k$  is set to arise out of under assumption you said to arise out of a stationary distribution. Stationarity is fundamental to anything that we can do computationally that is the fundamental key. So, it is the limitation that that is imposed are not because we wanted, but because if it is not stationary there are too many things we know how to do analytical. So, stationarity assumption is fundamental in almost all of the time series analysis. And likewise here I am assuming that I am having a parallel time series at each of the  $m$  locations, one for each stations. From each of these parallel time series I can now compute the statistics the individual mean, the mean vector, the individual variances, the covariances, these two together summarize the vector  $\bar{Z}$  and the matrix  $C$ , and that can be done if you give me a long series of time series at various stations.

So, we assume that we have availability of  $\bar{Z}$  and  $C$  at  $m$  stations, so that is what exactly is I am what I described is now reinforced in the slide. Let me quickly reinforce that. So, computational the spatial covariance  $C$  is the computation of spatial covariance among  $m$  observation stations.  $Z$  is the time series of the field variable,  $\bar{Z}$  is the vector of means. So, I am now doing instead of doing individual stations, I am collectively doing for all stations  $Z_k$  is the vector,  $\bar{Z}$  is a vector, I am trying to take

the average of capital T number of observations.  $\tilde{Z}_k$  is the vector of anomalies earlier I talked about the individual anomaly, now I can collect them in the vector. So, it is a vector anomaly.

If the vector of anomaly is known, I can now compute the covariance matrix as  $\tilde{Z}$  times  $\tilde{Z}^T$  the summation over  $k = 1$  over  $k$ . This C is symmetric positive definite that is the assumption. Why, if the number of samples is large, if the number of samples has been collected over a long period of time, I think one can readily see this covariance the covariance is always symmetric. If the number of data is small it may not be possible definite; if the number of data is large that the reason to believe that C so computed would indeed be symmetric and positive definite as well. So, I am going to make an assumption that C is symmetric and positive with given.

So, what the C captures that is important thing that is the fundamental idea of both Kolmogorov and Wiener; and they propose this in nineteen early 41-42. The matrix C captures the natural variation in the observation natural spatial variations. Please remember this is not related to the observation error covariance, I should have said error here, observational error covariance. Observational error covariance is R; in this case, R is zero because we have assumed the observations are perfect. So, this spatial covariance essentially captures the natural spatial variability of the field variable of interest in the region covered by in the geographical region covered by the m observation stations that is the fundamental idea, that is a starting point.

Now, let me go back and talk about why are we interested in this correlation. Please understand our ultimate aim is to predict; to predict, I need to assimilate. To predict there are only two kinds of things you can bank onto be able to generate a prediction one is causality, another is correlation. So, if you are trying to use models to predict, the models represent the causal relation that exists and that helps you describe the underlying physical process. For example, the dynamical system that relates to the motion of the earth around the sun the dynamical system that relates to the motion of the moon around the earth. This combined dynamical system we have understood very well and using it we are able to predict the lunar and solar eclipses to very high degree of precision. So, this dynamical system essentially kept has captured the causality principle that underlie the motion of the moon around the earth and the earth around the sun.



So, likewise every model be it static, be it dynamic, be it realistic, be it stochastic, all models in some sense encapsulate some form of this causality principle, momentum is conserved, energy is conserved, these are all causality principle. In the case of time series, I derive empirical models autoregressive moving average models they are empirical models. These empirical models in time series as essentially derived out of correlation. So, what is that we do we look at a time series, you compute this temporal correlation. We see how the temporal correlation decays with time. If it decays fast, the process has less memory; if it decays very slowly, the process has long memory. Mathematicians have already analyzed properties of several different types of time series models and they have catalogued and have created an album if you wish of correlation structures of various models.

So, given a time series data, you compute the correlation structure of the given time series and compare it against the album, and look at which one looks closest, it may not be one, it could be a subset of two or three. You narrow down the window, and then try to use each one of these models to be able to distinguish which model is better. So, what is the underlying theme to be able to predict either you need causality or correlation, this approach to optimal interpolation rests on our ability to predict based on correlation.

So, what does it mean if two quantity especially separator is positively correlated means what, if the variable in one of the stations increase, there is a likelihood the other station will also increase because there is a positive correlation. If there is a negative correlated, if one increase of other will decrease. So, I can infer from the increase or decrease of a particular quantity in a given station and knowing the correlation, I should be able to predict what will happen at the other station that is the fundamental principle that underlie predictive science based on correlation.

So, fundamentally ultimately the aim of data assimilation is to predict, data assimilation essentially tries to help you to fit the model to data fitting model to data is essentially model calibration. So, the forecast generate from calibrated models are better than the one generated from uncalibrated model that is why we do data assimilation. The alternative to that is to be able to understand spatial and temporal correlation structure, Wiener's theory applies both to temporal correlation analysis as involved in time series analysis as well as spatial correlation analysis as would be of interest in any geophysical

science. So, the matrix  $C$  captures the natural variability of the variable of interest in the chosen geographical domain.

(Refer Slide Time: 25:08)

### CROSS COVARIANCE BETWEEN OBSERVATION AND GRID

- Let  $\bar{x} = E(x)$  be the long time average of the field variables on the grid
- Let  $\tilde{x} = x - \bar{x}$  and  $E(\tilde{x}) = 0$
- Define  $D = E[\tilde{x}\tilde{z}^T] \in \mathbb{R}^{n \times m}$  – cross covariance between the grid values and the observations

Handwritten notes in red:

$$x \leftarrow z$$

$$\bar{x} \in \mathbb{R}^n \quad \tilde{z} \in \mathbb{R}^m$$

$$D = E[\tilde{x}\tilde{z}^T] \in \mathbb{R}^{n \times m}$$

Now, so I have please go back to my original picture here. I have a grid; I have an observation stations. So, the  $C$  is the covariance among the observation stations, but my interest are in computational grid. For example, if I have established an observation network, the observation network is going to be fixed in spatial distribution. Computational grid depending on the computing power, I can change the computational grid as well. If I have a larger computing power, I can have a smaller grid size and a larger grid number. If I have a smaller computing power, I may have a closed grid and a smaller number of total numbers of grid point. So, computational grid there is no fixed value for that, it depends on what else I am interested what kind of processors I am interested in analysis and so on.

So, now what is that we would like to be able to do, we would like to be able to transfer the knowledge from observation network to the computational network that has been the theme in the successive error correction, we have been talking about in the last lecture as well iterative methods. So, the theme is very similar except that this is this new idea is rooted deeply in the correlation structure. So, if I have values in the observation network, please recall our ability to transfer information from observation network to the grid network. This is  $R_m$ , this is  $R_m$ , this is  $Z$ , this is  $H$ , this is  $H$  transpose, this is  $H$

transpose. So, I can go from one network to another network by this interpolation scheme.

So, what is that we have done, from the observation stations, we can also think of having a time series of values at the grid point using this interpolation. So, if I have a time series, so let us fix a particular time noon January 1st, I have  $m$ -stations I have  $h$  matrix that can interpolate between the computational grid and the observation network. Yes, you may say hey if you interpolate it is not the interpolation is incur error, yes, it may I am cognizant of that, but I want the reader to appreciate that I have the ability to lift the information from one network with other network through  $H$  and  $H$  transpose.

So, let us pretend now I have a corresponding time series of the same physical very field variable of interest at each one of the  $n$  grid points the time series the is a saying. In other words at noon I have an observation station at noon I have interpolated value along the grid noon of today 2 o'clock of today, 5 o'clock of today every day. So, by hour by hour, I have a time series going over let us say 50 years of day, I can do that. Again I want you to recognize, while the observation network is fixed the computational grid may be changing. So, if I fix the computational grid and if the computational get embeds the observation network, there is one way to be able to lift it from the observation network to the computational grid.

So, once I have the same time series at the computational grid, I have a vector  $x$  from the vector  $Z$ . This vector  $x$  belongs to  $\mathbb{R}^n$ , vector  $Z$  belongs to  $\mathbb{R}^m$ . If I have a time series on the grid, I can compute its expected value then I can compute the anomaly. The anomaly is such that its mean is 0. Now, comes the important thing. Let  $D$  be a matrix that captures the cross correlation between the grid and the observation network. So,  $\tilde{x}$  is a vector that is defined on the grid, so  $\tilde{x}$  is a vector belonging to  $\mathbb{R}^n$ ;  $\tilde{Z}$  is a vector belonging to  $\mathbb{R}^m$ . So, if I multiply  $\tilde{x}$  by  $\tilde{Z}^T$ , you can think of this as this is the column that the row there is an outer product matrix, this matrix is going to be  $\mathbb{R}^{n \times m}$ .

If I took the expectation of this outer product matrix that is  $D$ , so what is  $D$ ,  $D$  is the  $n$  by  $m$  matrix that relates to the cross correlation between the grid variables and the observation variables. So, this is the cross covariance between grid values and the observation values variables. So,  $C$  and  $D$  are going to play a very major role in our

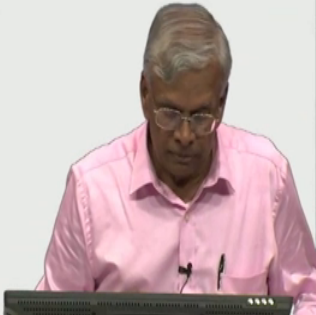
analysis. Now, please remember  $C$  may be fixed in time because observation network once we establish the network we cannot change, but  $D$  can change in time, I am sorry the comp computational grid can change in time,  $n$  can change in time therefore,  $D$  can change in time.

So, how do you get from observation to  $C$  to  $D$  is to be able to elevate interpolate from observation at the grid network this interpolation scheme, there are several such schemes I have already talked about a bilinear interpolation that interpolates between grid and the observation network. So, we can use one of these. So, what is that we have I have access to  $C$ , now I have access to  $D$ .  $C$  is the covariance of the given field variable of interest at the observation location;  $D$  is the cross covariance between the grid and the observation network. So, you have there is a lot of statistical computations. So, if you have a time series over a 100 years, if you have a time series over 100 years hour by hour, minute by minute, your data set may be very large from that set data we have to crunch. The value of the matrix  $C$ , you have to crunch the value of the matrix  $D$  that could be done; very routine calculation it does not take too much time.

(Refer Slide Time: 32:26)

### COMPUTATION OF CROSS - COVARIANCE

- Let  $x_k = (x_{k,1}, x_{k,2}, \dots, x_{k,n})^T \in \mathbb{R}^n$  be the vector values of the same field variables at the grid locations at time  $k$
- Given  $\{x_k \mid 1 \leq k \leq T\}$ , the time series
- Then  $\bar{x} = \frac{1}{T} \sum_{k=1}^T x_k$ , sample average  $\rightarrow$  (4)
- Let  $\tilde{x}_k = x_k - \bar{x}$ , anomaly  $\rightarrow$  (5)
- Then  $D = \frac{1}{T} \sum_{k=1}^T \tilde{x}_k \tilde{x}_k^T \in \mathbb{R}^{n \times n} \rightarrow$  (6)



Here I am just trying to describe formally the notion of computing the elements of the matrix; I want to reinforce that again. So, let us quick quickly run through these calculations. Let  $x_k$  be the state variable vector of the same field variable at the grid locations at time  $k$ , we may have at times. So, if  $Z$  has a time series that the

corresponding time series for  $x$ , once I have the time series I can compute the mean, I can compute the anomaly, I can compute the cross product which is  $D$  that is what. So, the there we talked about a concept here we talked about algorithms. So, concepts algorithms concepts algorithms algorithm for  $C$ , algorithm for  $D$  are given

(Refer Slide Time: 33:20)

### STATEMENT OF THE PROBLEM

---

- Let the observational covariance matrix  $C$  and the cross covariance matrix  $D$  be known
- Given a new observation vector  $Z$ , how to optimally compute the induced grid values  $x$
- Let  $\tilde{Z} = Z - \bar{Z}$  anomaly of  $Z$  w.r.to the known long term average  $\bar{Z}$
- Similarly,  $\tilde{x} = x - \bar{x}$
- Background  $\bar{x}$  is known, we can recover  $x$  once  $\tilde{x}$  is computed

Now, what is the statement of the problem? Let I am sorry there is a spelling problem here that has to be it. Let the observational covariance  $C$  and the cross covariance  $D$  be known, matrix  $C$  and matrix  $D$  be known. So, what is  $C$  and  $D$  represent  $C$  and  $D$  represents the stationary values of the covariance between observations and covariance and the cross covariance between observation the grid because the underlying process of stationary,  $C$  and  $D$  does not change in time especially  $C$  and  $D$  does not change in time if you consider a long time series.

So, what stationarity means I want to comment on that little bit. If the regime of the climate has changed, the underlying distribution would have changed. So, what the stationarity assumption means in practice, we are assuming the regime under which that climate is operating has not changed statistical that is the idea. So, how do you know the regime we are operating has changed or not that is a different question. We should have a long time series, you should break the time series into different parts, we should calculate different quantities in different times and see whether these statistical quantities has

changed over a 100 years, first hundred year, second hundred year, third hundred year, first fifty year, second fifty year, first ten years, second ten year.

So, decadal variation, essentially variation, annual variation, so one can do statistical tests for regime changes the regime shifts, if one has access to a reasonably long time series. In other words, one can test the hypothesis namely if the underlying statistics is invariant or changed. So, it all depends on how much data you have. If you do not have much data, you cannot even do that testing. If you do not have much data, if you cannot do that testing, just pray the lord and assume stationary. So, that is these are some of the key things that one has to keep in one's mind now.

So, what does C and D refers in some sense they refer the climatology and that important thing here. C and D refers to the underlying climatology, the stationary aspects of the climatology. Suppose a New Year dawns January 1st 2016 today. I have computed C and D over the past years. A new observation from my observation network arises on this new day. These Z observations available only the observation network, but I would like to optimally compute the values induced on the grid  $x$ .

So, I hope you understand what I am talking to talking about. C and D are known, C and D are based on past values. They represent climatology a new day dawns. A new day brings a new observation Z. The new observation C is confined to observation location  $m$  of them, but I would like to be able to elevate that  $m$  to  $n$  grid points how do you do that how do you do it optimally. So that is the step into the problem given a new observations Z and the climatological information in embedded in C and D, how to optimally compute the induced grid value  $x$  from Z.

So, I know  $\bar{Z}$  what is the  $\bar{Z}$  is a long term average of Z because stationarity I assume  $\bar{Z}$  has stabilized. So,  $\bar{Z}$  is invariant in time. So,  $\tilde{Z}$  is equal to  $Z$  minus  $\bar{Z}$ , there is anomaly of Z with respect to the long term average  $\bar{Z}$ . Similarly let  $\tilde{x}$  be  $x$  minus  $\bar{x}$ , I know  $\bar{x}$  because I have computed D with it. I do not know  $x$ ,  $x$  is the one I want to be able to determine, but instead of computing  $x$ , I am going to compute  $\tilde{x}$ . So, computing  $x$  is equal to computing  $\tilde{x}$ , because if I know  $\tilde{x}$  I can add  $\tilde{x}$  to  $\bar{x}$  to get  $x$ . So, I am going to work with anomalies.

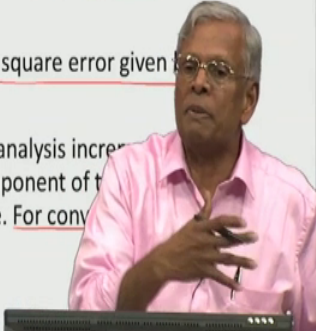
To at  $\bar{Z}$  be the observation anomaly on a new day, let  $\bar{x}$  I am sorry that is a  $\bar{x}$  I am sorry  $\tilde{Z}$  be the observation anomaly on a given new day. Let  $\tilde{x}$  be the

corresponding induced anomaly in the grid, my job is to be able to optimally determine  $\tilde{x}$  from  $\tilde{Z}$  knowing  $C$  and  $D$ . Please understand the background  $\bar{x}$  is known, we can recover  $x$  once  $\tilde{x}$  is computed. So, we are simply going to concentrate on computing the  $\tilde{x}$  anomaly.

(Refer Slide Time: 38:57)

### OI APPROACH

- The basic idea is to express the analysis increment  $\tilde{x}_i$  as a linear combination of the observation increment  $\tilde{Z}$
- Let  $w = (w_1, w_2, \dots, w_m)^T \in \mathbb{R}^m$  be the unknown weight vector
- Let  $\tilde{x}_i = \sum_{j=1}^m w_j \tilde{Z}_j = w^T \tilde{Z} \rightarrow (7)$
- Find the  $w \in \mathbb{R}^m$  that minimizes the mean square error given by
 
$$f(w) = E[\tilde{x}_i - w^T \tilde{Z}]^2 \rightarrow (8)$$
- Note: Here  $\tilde{x}_i$  is the  $i^{\text{th}}$  component of the analysis increment vector  $\tilde{x} \in \mathbb{R}^n$  at a given time and  $\tilde{Z}_i$  is the  $i^{\text{th}}$  component of the observation increment vector  $\tilde{Z} \in \mathbb{R}^m$  at the same time. For convenience, the index is suppressed



So, what is the optimal interpolation approach please understand this is one of the earliest known methods in predictive science. The basic idea of this optimal interpolation is to express the analysis increment. What is the analysis increment,  $\tilde{x}$  minus  $\bar{x}$ . So,  $\tilde{x}_i$  what is  $\tilde{x}_i$ ,  $\tilde{x}_i$  is the new increment at the location at the grid point  $i$  that is to be gained from the new observation. I am assuming I can express my  $\tilde{x}_i$  as the linear combinations of the observation increments  $\tilde{Z}$ .

So, let me I want to make sure my  $\tilde{Z}$  is not that is right they are  $\tilde{Z}$ . So,  $\tilde{Z}$  is the vector of observation increments. I am going to compute location by location grid location by grid location  $i$  is the  $i^{\text{th}}$  grid location. The  $\tilde{x}_i$  is what I want to compute now I have to concoct a model that relates the known to the unknowns  $\tilde{x}_i$  is not known  $\tilde{Z}$  is known. So, I am going to confine my attention to the class of recovery process to a class of estimators, where  $\tilde{x}_i$  is expressed simply as a linear combinations of the elements of  $\tilde{Z}$ . Are you with me? So, I want you to think about that.

So, what is that we are going to do, if I am going to have a linear combination I need to have the weights. So, let  $w$  be the vector of weights that is going to be using the linear combination  $w$ 's are unknown weight vector. So, given this philosophy of connecting the known to the unknown, let the unknown be the linear combination of the known. So, I could express this as  $w^T \tilde{Z}$ , therefore under what condition  $\tilde{x}_i$  will be optimal, it will be optimal under optimal choice of  $w$  vector. So, the whole problem now reduces to finding an optimal vector optimal weight vector  $w$ . Optimal in what sense again we come back to least squares, both Kolmogorov and Wiener independently in 1941, they were separated geographically they did not talk to each other, but they had come up with very similar ideas.

So, the mathematical problem now reduces to the following find a vector  $w$  belonging to  $\mathbb{R}^m$  that minimizes the mean square error between the value of the increment at the  $i$ th the grid from the linear combinations because underlying quantities are all naturally random I have to take the expected value. So, the objective function here is the mean of this square of the error. So, excise that  $i$ th component analysis increment vector at a given time. So, all these things are done in a given time. If you are interested in various time I had repeated at a given time. So, with a loss of generality I am have assume  $k$  is fixed  $k$  the time is fixed. So,  $\tilde{Z}$  is the vector,  $\tilde{Z}_i$  is the  $i$ th component of the observation increment for convenience we have suppressed the time index because we have fixed that time this is an analysis being done at a given time. So, if you have a subroutine that can do this, I can repeatedly use this for various times.



(Refer Slide Time: 43:16)

## OI APPROACH

- Rewrite:
 
$$\begin{aligned}
 f(w) &= E[\tilde{x}_i^2 - 2\tilde{x}_i(w^T \tilde{Z}) + (w^T \tilde{Z})^2] \\
 &= E(\tilde{x}_i^2) - 2E[\tilde{x}_i \tilde{Z}^T]w + w^T E[\tilde{Z} \tilde{Z}^T]w \\
 &= \text{Var}(\tilde{x}_i) - 2d_{i*}w + w^T Cw \rightarrow (9)
 \end{aligned}$$

$E[\tilde{x}_i \tilde{Z}^T] = i^{\text{th}} \text{ row of } D$   
 $= d_{i*}$

where  $E[\tilde{x}_i \tilde{Z}^T] = d_{i*} = i^{\text{th}} \text{ row of } D$  inner product  
 $E[\tilde{Z} \tilde{Z}^T] = C$ , observation Covariance

11

I hope the formulation of the problem is clear now. Look at this now. I am falling back to least squares, but I am falling back to least squares not with respect to static model or dynamic model. But based on correlation structure, cross correlation structure, stationarity assumption and so on.  $f$  of  $w$  can be now written as this because is like a minus b whole square expectation operator I can pull it into each of the term expectation of the sum is some of the expectations. The first term is the variance of  $x_i$ 's expected expectation of the square of  $x_i$  tilde is the variance. Expected value of  $x_i$  tilde  $Z$  tilde transpose it is essentially the  $i$ th row of  $D$ . So, that is denoted by  $D_i^*$  please remember  $D$  is the cross covariance between the grid and the observation location. So, I know every grid point will be related to every observation locations.

So, here I would like to talk about a particular difference between Crosman scheme another scheme. In the Crosman scheme, what is that we assume give a point, the number of observation station that I can affect the given grid point are those that are lying within the radius of influence  $D$  that came in 1950s mid 50s. In Wiener's time, in nineteen early 40s, 41-42, they did not restrict the influence of one and the other. They assumed everybody is going to influence everybody, but they interest they were measuring the influence through correlation and cross correlation. So,  $i$ th grid point has a cross correlation with every observation station and that is given by the  $i$ th row  $d_i^*$ . So, the this now becomes 2 times  $d_i^* w$ .

Now I would like to go back to the  $w$  is a column vector I want you to understand  $w$  is a column vector  $d_i$  is a row vector. So,  $d_i^T w$  I am sorry  $d_i^T w$  that is an inner product, it is a scalar because one is the row another is a column vector. Now, the last term is essentially  $w^T C w$  you remember  $C$  the correlation among observation locations. Now, look at this equation 9. Equation 9 relates to the  $i$ th row vector of  $D$  then entire matrix  $C$  and the variance of the field variable at the  $i$ th grid point all of them are known what is the only thing that is not known  $w$ . Now, please also realize this  $f(w)$  in 9 is a quadratic form. So, what is the mathematical problem pick  $w$  such that this quadratic form is the minimum how many times we have minimized quadratic form in this class million times. So, one is quadratic variable, another linear variable another is the constant.

(Refer Slide Time: 47:16)

### OI - ALGORITHM

---

- $\nabla f(w) = -2d_i^T + 2Cw \rightarrow (10)$
- $\nabla^2 f(w) = 2C, \text{ SPD} \rightarrow (11)$
- Setting the gradient to zero, optimal  $w$  is the solution of  
 $Cw = d_i^T \rightarrow (12)$
- By repeating this process for each  $1 \leq i \leq n$ , we can obtain the analysis increment across the grid
- Since the  $n$  systems, one for each  $i$ , has the same matrix  $C$  on the l.h.s, we can solve it efficiently using the cholesky decomposition of  $C$

$C \begin{matrix} m \times m \end{matrix} w \begin{matrix} m \times 1 \end{matrix} = d_i^T \begin{matrix} 1 \times m \end{matrix}$   
 $A \begin{matrix} n \times m \end{matrix} x \begin{matrix} m \times 1 \end{matrix} = b \begin{matrix} n \times 1 \end{matrix}$

So, computing the gradient  $f$  of  $f(w)$  in 9, I am not going to do the arithmetic, I very strongly encourage you to apply the principles from multivariate calculus that we have talked about. The gradient is given by 10; the hessian is given by  $2C$  we have already assumed  $C$  is SPD. So, hessian is SPD. So, if I equate the gradient to 0 given that the hessian is a SPD, it must be a minimum because the function quadric function is a convex function is a unique minimum. So, the solution of 10 is given by 12. Let us look at that now,  $C$  is the matrix,  $C$  is the  $m$  by  $m$  matrix;  $w$  is a vector which is  $m$  by one vector that must be equal to  $d_i^T$  transpose. Please remember  $d_i^T$  transpose is  $i$ th the row.

Let us go back now.  $D$  is a  $n$  by  $m$  matrix. Therefore, in the matrix  $D$  if I consider the  $i$ th row the  $i$ th row is the  $m$  vector because this is  $m$  therefore,  $d_i^*$  is an  $i$ th row  $d_i^*$  transposes the  $m$  column vector  $n$ . So, this is of the form  $Ax = b$ , where  $A$  is an SPD. Now, do you remember this is the linear system that symmetric positive definite matrix I want to solve this, how many different methods we have seen I can solve this by Choleskey, I can solve this by QR, I can solve this by SVD, I can solve this by iterative scheme such as Gauss-Seidel that ever so many methods. I can also solve this by gradient method; I can also solve this by conjugate gradient method.

Now, you can see the power of the mathematical tools that we have already developed that is the key we are looking at the foundational aspects of algorithms to do data assimilation, no matter what the variety of assimilation process we are involved in. Please remember this already idea is extremely different from any other ones we have already seen yet the tools may have developed are very helpful computing the quadratic form, compute the gradient hessian minimizing the quadratic form, solving the resulting equation by matrix methods by direct optimization methods those are all tools in your toolbox. In fact, I would like to remind all of you that the famous saying if hammer is the tool, if hammer is the only tool you have every job looks like a nail, I want all of you to write this down and put it in your study.

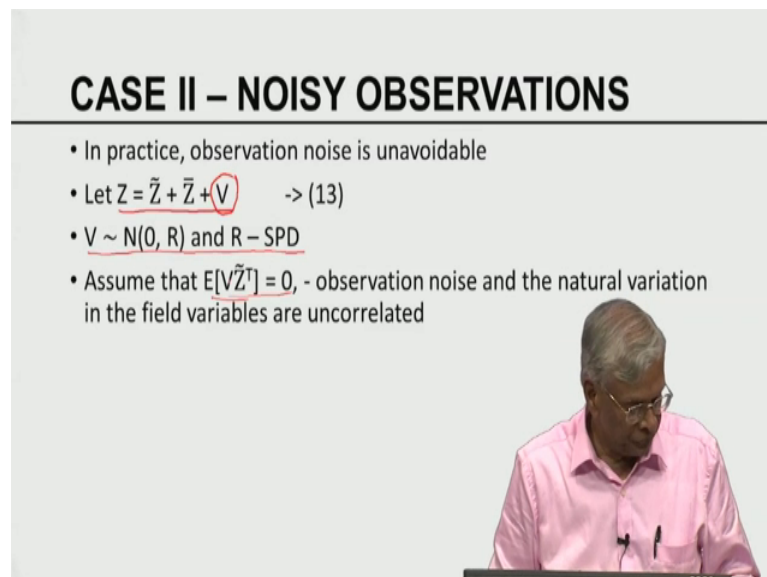
If you have only one tool and that too happens to be hammer what is that you can do the hammer you can only hit. So, the famous proverb is if hammer is the only tool you have every job looks like a nail, you can really hit the nail. So, what cannot do depends on what you do not know what you do not have. So, larger the toolbox is better days, now please understand our toolbox has been filled with tools which tools that can be used in solving problems. Therefore, 12 is the solution of a symmetric positive definite system, I can solve this by one of several methods. I would like the reader to indulge in this process and convince oneself that I can solve twelve by one of many methods. Now, by repeating this process for each  $i$  we can obtain analysis increment over the entire grid.

But, now look at this now, if I want to solve this, so for every  $i$ , the left hand side remains the same only the right hand side changes, are you with me. So, we are going to have a collection of  $n$  linear system with the same matrix  $C$ , but the right hand side is different. What does it mean if you compute the cholesky factor for  $C$  once I can use it repeatedly for every solution for various grid points, so that is a computational

advantage. So, we can solve this efficiently by using cholesky decomposition of  $C$  once and repeatedly using the cholesky factor to solve the grid variables at each of the  $n$  locations.

This method has come to be called this method has come to be called optimal interpolation. In the area of geophysics, there is a method called Kriging, I am sure many of you who are in geophysical sciences could have heard of that Kriging is a method essentially an optimal interpolation method. So, Wiener's method have been applied to many, many different fields of activity it has been applied under the camouflage of very, very different names Kriging is one such name that has been applied in geology and geophysics. Kriging is if you look at the mathematics of it, Kriging is essentially Wiener Kolmogorov optimal interpolation the person who popularized the application of optimal interpolation especially in atmospheric as well as oceanographic sciences is the Russian scientist Lev Gandin. Gandin has published a book that is totally devoted to application of OI to solve several different problems of interest in climatology oceanography atmospheric sciences and so on, and I would strongly recommend the reader to take a look at this extremely good book by Lev Gandin.

(Refer Slide Time: 53:07)



**CASE II – NOISY OBSERVATIONS**

- In practice, observation noise is unavoidable
- Let  $Z = \tilde{Z} + \bar{Z} + V$   $\rightarrow$  (13)
- $V \sim N(0, R)$  and  $R$  – SPD
- Assume that  $E[V\tilde{Z}^T] = 0$ , - observation noise and the natural variation in the field variables are uncorrelated

The slide includes a video inset in the bottom right corner showing a man with glasses and a pink shirt speaking at a podium.

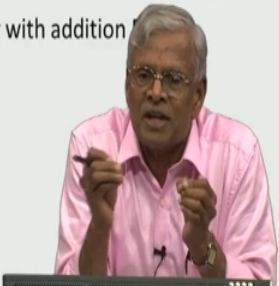
Now, I am going to talk very quickly about the extensions; case two noisy observations in practice observations are noisy noise is unavoidable. Therefore, my observation  $Z$  is equal to  $Z$  bar is the mean,  $Z$  tilde as the anomaly until now we had only  $Z$  is equal to  $Z$

bar plus  $Z$  tilde. Now, we have a new guy coming into the table, coming to the game and that is  $V$ , I am going to assume is Gaussian,  $R$  is SPD. I am also going to assume that the natural variability in  $Z$  and the observation noise are in correlated, that makes sense? The climate does not care for the instrument you use to measure. So,  $V$  essentially comes from the measuring process the measuring process is not or it does not affect the natural variability of a climatic system. So, the cross correlation between the natural variability of field variable interest and the observations are zero.

(Refer Slide Time: 54:17)

### COVARIANCE OF OBSERVATIONS

- $\text{Cov}(\tilde{Z}) = E\{[(Z - \bar{Z}) - V][(Z - \bar{Z}) - V]^T$   
 $= E[(\tilde{Z} - V)(\tilde{Z} - V)^T]$   
 $= E[\tilde{Z}\tilde{Z}^T] + E(VV^T)$   
 $= \underline{C + R} \quad \rightarrow (14)$
- Observational Covariance becomes larger with addition

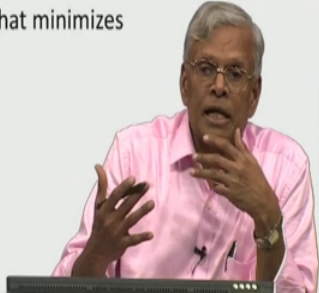


In this case, I can now compute the covariance of since we have discussed many of these things in detail I am going to hit all the major points. So,  $Z$  tilde is given by this, so its covariance is  $Z$  tilde times  $Z$  delta transpose. There are four terms the cross terms vanish because there is no covariance between  $Z$  and  $Z$  tilde. So, we are left with covariance of  $Z$  tilde is  $C$  plus  $R$ . Now, look at this now what is the role of  $R$  increases the covariance; earlier  $R$  was zero because there is no observation noise observations are perfect this is the only difference rest of it all the same. So, wherever there are  $C$ , you replace  $C$  by  $C$  plus  $R$  solve the problem.

(Refer Slide Time: 55:06)

### OI - APPROACH

- It is reasonable to assume that the analysis increment  $\tilde{x}_i$   $1 \leq i \leq n$  is uncorrelated with the observation noise  $V_j$ ,  $1 \leq j \leq m$
- $E[\tilde{x}_i V_j] = 0$ ,  $1 \leq i \leq n$ ,  $1 \leq j \leq m$
- $D = E[\tilde{x}(\tilde{Z} - V)^T] = E[\tilde{x}\tilde{Z}^T]$
- Given  $Z - \tilde{Z} = \tilde{Z} - V$ , find a vector  $w \in R^m$  that minimizes  
 $f(w) = E\{[\tilde{x}_i - w^T(\tilde{Z} - V)]^2\}$



So, again reasonable to assume the analysis increment is correlated with it, it is reasonable to assume that the analysis increments is uncorrelated with observation noise sorry, it is uncorrelated and that is described by this. So, I can define D, D is given by the similar formula. I have an f w, which is the weight factor. So, the whole thing I am simply again assuming the grid analysis increment is the linear combination of the observation analysis increment now, how good this is the linearity assumption you can change that, but just get the ball started Kolmogorov and the Wiener assumed that the given increment can be expressed as a linear combination at the observed one.

So, you can probably in a given situation, you can try other combinations of the dependence of  $\tilde{x}$  and  $\tilde{Z}$  and do the analysis, but the trouble is if the instantly bring in non-linearity that is going to computational trouble. So, you have to worry about is the trouble worth it; if you take the trouble and spend more money to solve the non-linear problem is it going to improve your prediction estimates. So, these are all open questions, I am not going to indulge in all other possibilities, it is simply to provide you a new way of thinking about prediction.

(Refer Slide Time: 56:47)

## OI - ALGORITHM

- $f(w) = E[\tilde{x}_i^2] - 2E[\tilde{x}_i(w^T(\tilde{Z} - V)) + E[w^T(\tilde{Z} - V)(\tilde{Z} - V)^T w]$   
 $= \text{Var}(\tilde{x}_i) - 2E[\tilde{x}_i(\tilde{Z} - V)^T]w + w^T E[(\tilde{Z} - V)(\tilde{Z} - V)^T]w$
- $\nabla f(w) = -2d_{i*} + 2(C + R)w = 0$
- $\nabla^2 f(w) = 2(C + R)$
- Optimal  $w$ :  $(C + R)w = d_{i*}^T \quad 1 \leq i \leq n$
- Note: Knowing the correlation structure of the field variables, ~~w~~ We can easily lift the information from the observation network to the computational grid

16

So, if you consider  $f(w)$  and again I given all the details I would leave this as an exercise. So, this is the quadratic form. You compute the gradient, you compute the hessian, you can see wherever there were  $C$  is replaced by  $C$  plus  $R$ . So, except for that difference that is not much. So, the optimal  $w$  is now obtained as a solution of this. If you set  $R$  is equal to 0, I get the previous equation. So, what is the fundamental idea knowing the correlation structure of the field variable we can comma, this is lower case, we can easily lift the information from the observation network to the computational grid that is the that is the essence of this discussion.

(Refer Slide Time: 57:37)

## REFERENCES

- This Module follows Chapter 19 in LLD (2006)

17

This module is contained in chapter 19 that where we have discussed this at greater length, book by Lewis Lakshminarayanan (Refer Time: 57:45) 2006. With this I hope we have given you an opportunity to develop an appreciation for and another way of thinking about prediction, prediction creating analysis, how to transfer the new information and combine it to the old information. I want to talk about that part a little bit  $Z$  the new observation that is a new information,  $C$  and  $D$  old information. So, in some sense you can think  $C$  and  $D$  as an equivalent of prior,  $Z$  has a new information maybe basis tripping in very quickly here. There are two pieces of information I am trying to combine them. I am trying to combine them so as to minimize the error in the estimate least squares comes in creeps in secretly. So, base comes in secretly, least squares comes in explicitly, you can see the commonality between various techniques even though the basic assumptions with which these techniques are developed are totally different.

With that, we conclude our discussion on the optimal interpolation techniques.

Thank you.