SUPPLEMENTARY LECTURE NOTES FOR

ATOC 7500

MESOSCALE ATMOSPHERIC MODELING

SPRING 2008

Author

Tomi Vukicevic

CHAPTER 1 - DRAFT

Inverse Modeling and Data Assimilation in Geosciences: A Practial Guide

Author : Dr. Tomislava Vukicevic (*tomislava.vukicevic@colorado.edu*)

Affiliation : Associate Professor, Department of Atmospheric and Oceanic Sciences (ATOC), University of Colorado, Boulder, Colorado, USA

Date of this document: December 2006

Introduction

1. Why inverse modeling and data assimilation in the Earth System sciences?

This question could be answered in two ways depending on whether one is starting form the point of view of modeling or the need to produce quasi-continuous environmental data on a spatio-temporal grid. Both views are presented in the following.

1.1 Modeling in the Earth System Sciences

Physical theories in the Earth System sciences are designed to explain and possibly predict natural phenomena. The explanation by a theory is also a form of prediction as it states certain consequences for certain causes. Both, the explanation and prediction typically include quantitative representation of the natural system state.

Quantitative assessment of the actual, true, state is fundamentally achievable only by measurements. The Geophysical theories are consequently designed to explain and predict the measurements.

The theories are most often expressed in a form mathematical relationships which define a model. The model in general represents governing physical laws and includes a set of quantities which entirely define the state by the model. The defining quantities are called control parameters. The control parameters are initial and boundary conditions, external forces and other physical quantities which define medium or environment for a process that is modeled.

The quantification of the system state by application of an assumed set of control parameters by the governing laws is called ***forward model*** of the system state or simulation of the measurements. Obviously, under conditions of well known governing laws and accurate quantification of the control parameters the forward model would produce accurate simulation of the measurements and would have ability to predict future states. It is common, however, that the governing physical laws are known but the control parameters values are not. This condition occurs in variety of models which are based on application of fundamental laws for macro scale phenomena such as conservation of energy and mass, propagation of energy through media and bulk energy and mass transformations. Examples of this type of model are found across the Geoscience disciplines such as in the Atmospheric Sciences, Oceanography, Biogeochemistry and Hydrology. What is typically not well known are initial and boundary conditions, some external forcing mechanisms and bulk properties of the medium in parameterized energy and mass transformations.

Because the model simulates the measurements, it is natural to ask whether there is a formal and objective way to use the measurements to infer the correct control parameter values for the model? This problem is called ***inverse problem*** or ***inverse modeling problem***. When there is a need to use the measurements to infer the control parameters frequently in a quasi continuous manner over time the inverse modeling problem is often referred to as ***data assimilation***. Thus, the first way of answering the question "Why inverse modeling and data assimilation in the Earth System sciences?" is

To objectively correct modeled state of the system or a component of it by using measurements, such that the model could be effectively used to analyze and predict the system.

The concept of inverse modeling and data assimilation for the purpose of improving model simulation and prediction has been used first in physical sciences in 17-th century in works by Euler, Lagrange and Laplace on calculating orbits of celestial bodies (Lewis et al, 2006). Gauss first formally described a method of data assimilation in his book on "Theoria Modus Corporum Coelestium" written in 1809. The data assimilation approach by Gauss is summarized in the following quote from the book:

" If astronomical observations and other quantities on which the computation of orbits is based were absolutely correct, the elements aso, wheter deduced from three or four observations, would be strictly accurate, so for indeed as the motion is supposed to take place exactly according to the laws of Kepler, and, therefore, if other observations are were used, they might be confirmed but not corrected. But since all our observations and measurements are nothing more than approximations to the truth, the same must be true of all calculations resting on them, and the highest aim of all computations made concerning concrete phenomena must be approximate, as nearly as practicable, to the truth. But this can be accomplished in no other way than by suitable combination of more observations than the number absolutely required for the determination of the unknown

quantities. This problem can only be properly undertaken when an approximate knowledge of the orbit has been already attained, which is afterwards to be corrected so as to satisfy all the observations in the most accurate manner possible."

In the chapter 3 we come back to the Gauss description of the data assimilation problem as it includes not only statement of the fundamental purpose of solving the problem but also key properties of the measurements and model. Since mid 19-th century the inverse modeling and data assimilation methodology continued to develop mostly in technical engineering applications where the primary problem was to either optimize a controlled system performance or to devise inference of signal over noise for measurements of dynamical systems (Jazwinski, 1970).

In the Geosciences the data assimilation for improving the model simulations was first explicitly used in 1980-es in Numerical Weather Prediction (NWP), where there is direct need to improve initial conditions to improve forecast skill (Daley, 1990; Kalnay, 2004). More recently, new research is being performed in which other than initial condition parameters are improved by the inverse modeling and data assimilation methods (Braswell et. al, 2005; Vukicevic et al., 2000).

1.2 Environmental data by the data assimilation

The repeated data assimilation in the NWP applications naturally resulted in a time record of the Atmospheric states which have been used as data for other than weather forecast purposes. For example, the weather analysis data are used in physical analyses of short term weather phenomena as well as in climate analysis and research on the climate processes. In late 1990-es the data assimilation emerged as necessary

procedure in most other analysis of the Earth System states such as the analysis of oceans, land surface, soil, atmospheric trace gasses and particulates, land hydrology and biogeochemistry. Every one of these data analysis includes a model which control parameters are repeatedly corrected by the assimilation of measurements.

Examples of the Earth System data types containing temporally and spatially distributed physical quantities produced by some type of data assimilation are:

Ocean physical state data of
- velocity components
- pressure
- density
- temperature
- salinity

Ocean biological and chemical state data of
- concentration fields of nutrients
- plankton
- dissolved and particulate matter

Atmospheric physical state data of
- temperature
- pressure
- wind
- humidity
- cloud properties
- precipitation

Atmospheric chemical state and particulates data of
- concentration of trace gasses
- aerosols

Land surface state data of
- temperature
- moisture
- fluxes of gasses and energy
- optical and physical characteristics

Land biosphere data of

- land cover
- physiological characteristics of plants

Land hydrology data of

- hydraulic conductivity
- capillary pressure
- drainage coefficient in wetland areas
- leakage coefficient for river
- runoff

Soil state data of

- temperature
- moisture
- physical and chemical characteristics

These data are produced at operational environmental data centers and/or at sponsored national institutes and University research centers. In the USA the environmental data centers are sponsored or contained within all major national agencies: NOAA, NASA, DOE and DOD.

Thus, the second way of answering the question "Why inverse modeling and data assimilation in the Earth System sciences?" is

To produce quasi continuous fields of spatially and temporally distributed data of the Earth System, such that these data could be effectively used for the assessment of the system and for prediction.

CHAPTER 2 - DRAFT

Inverse Modeling and Data Assimilation in Geosciences: A Practial Guide

Author : Dr. Tomislava Vukicevic (*tomislava.vukicevic@colorado.edu*)

Affiliation : Associate Professor, Department of Atmospheric and Oceanic Sciences (ATOC), University of Colorado, Boulder, Colorado, USA

Date of this document: December 2006

---

2. Inverse problem basics

2.1 Parameter and model space

Let a model within scope of the Geoscience problems be denoted $\phi(m)$ with control parameters m, where m is a set of physical quantities. For each choice of the parameters, which is the quantitative specification of the elements of m, there is different realization or simulation with ϕ. The space or manifold spanned by the different values of m is called ***parameter space***. The parameter space is populated with possible values of the parameters. The existence of the parameter space (i.e., the many possible values) implies both the possibility of different parameter values for the modeled natural system and the stochastic nature of the parameter quantities. It is shown later in this chapter that the two causes for the existence of the parameter space cannot be easily distinguished in the inverse problem. The consequence of the parameter space is, however, the same. That is the quantitative result of $\phi(m)$ for different choices of the parameter values

renders a space or manifold, called the ***model space***. Because the fundamental property of model is to predict measurements, the existence of space of ϕ is interpreted as the existence of different possible simulations of measurements by the model which by design is intended to represent the same governing laws as the natural system that is measured. This intent is not necessarily always realized.

2.2 Measurement space

The measurement space is simpler to define than the parameter space. It is the space spanned by possible values of the measured quantity within the uncertainty range of the measuring procedure. The measuring procedure could include multiple measurements of the same quantity or multiple measurements of different quantities but for the same realization of the natural system. In the latter case the measurements constitute a multidimensional space similar to the control parameter set. In the oscillator model (1.1) there is only one dimensional phase space represented by χ. The measurable quantity in this example is $\chi(\tau_i)$, where τ_i is discrete time.

In the inverse problem there is explicitly derived dependency of the parameter to measurement uncertainty which is presented in the next section. Here the interest is to discuss the consequence of existence of the measurement space spanned by the measurement uncertainties. The range of control parameter values which would result from the measurement uncertainty is interpreted as in the forward problem as the range of uncertainty on the parameters. This property emphasizes the critical property of the modelization of the natural systems: ***When it is necessary to solve the inverse problem in***

the process of understanding and modeling of the natural system, the uncertainties in the measurements would render the uncertainties in estimates of what controls the system as hypothesized by the system model.

The parameter space which results from the variable external causes leading to the variable parameter values is related to the measurement space in more complex way than the measurement uncertainties. Each individual measurement is a recorded quantity of a response of an instrument to the medium that is measured. The medium when measured is at one specific state after one realization of the possible external cause. In order to capture natural variability of the parameters in the inverse problem solution which is caused by conditions external to the model, it is necessary to evaluate it from many measurements and different state realizations. It is shown later that validity of an evaluated range of actual variability in the inverse problem solution would depend on three factors: 1) abundance of measurements, 2) size of measurement errors and 3) strength of sensitivity of the forward model to the control parameters. Analysis of impact of each of these factors is important subject in specific applications as it addresses potential to distinguish different causes of the natural phenomena by the specific model and available measurements.

2.3Probabilistic nature of information in the inverse problem

The property of measurements to always have errors makes them random or stochastic quantities. Consequently, the model control parameters which would be

derived from the inverse problem solution using the measurements would also be random quantities. Even without the inverse problem the model control parameters could be random quantities if their values are uncertain. The random quantity, also called the stochastic quantity, is a quantity which exact value is not known or predictable. What is known about the random quantity is a possible value from a range with an associated probability. Because the measurements and control parameters are by design the stochastic quantities, relationships between these quantities in the inverse modeling problem and applications in the data assimilation problems must be derived based on the relationship between the associated probabilities.

2.3.1 Interpretation of probability

First, let A be realization of a stochastic physical quantity with the numerical value from within an interval $(x, x+dx)$. If there are many realizations of A, it would be possible to derive probability of A as chance of occurrence of A. A is then an event with probability $P(A)$ for which the following classical axioms of probability apply

$$\begin{aligned} &P(A) > 0 \\ &P(\emptyset) = 0 \end{aligned} \qquad (1.3)$$

If A and B are disjoint events

$$P(A \cup B) = P(A) + P(B) \qquad (1.4)$$

If A and B are not disjoint events

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad (1.5)$$

where $P(A \cap B)$ is joint probability

A distribution of probabilities over the space of possible values of A defines the probability distribution on that space. Another important function associated with the stochastic quantity and its probability distribution is the probability density $p(x)$ which satisfies

$$P(A) = \int_A p(x)dx \qquad (1.6)$$

where x represents coordinates, indicating that in the general case, the event A is a set of physical quantities included in A, such as the set of control parameters or set of measurements. The probability density function is of critical importance in the description of the stochastic quantities because when it is known they are completely described.

There is another intuitive way to interpret the probability of stochastic physical quantity in the inverse modeling problem. The probability could be defined as in Tarantola (2005) as : " *subjective degree of knowledge of the true value*". It is somewhat difficult to understand the emphasis on subjective knowledge in the Tranatola's definition, but it is instructive to consider the interpretation of probability which uses the reference to the truth. In this approach the uncertainty or error which renders the physical

quantities stochastic is measured as deviation from the truth. It is shown later that even when the truth is not known, which is most of the time, the uncertainty defined as deviation from the truth could be sensible approach to interpreting the probability in the results of the inverse modeling problems. The probabilistic variables and relationships (1.3 – 1.6) are the same for either interpretation of the probability.

2.4 General inverse problem and solution

2.4.1 Conditional probability

The key relationship which links the probabilities of stochastic quantities in the problem of evaluating the control parameters by inversion from the measurements is most commonly expressed by the Bayes' rule (1763) for conditional probabilities.

$$P(B/A) = \frac{P(A/B)P(B)}{P(A)} \tag{1.7}$$

where A and B are statistical events. The rule is actually derived from the definition of conditional probability

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \tag{1.8}$$

The left hand side is red as “probability of A given B”. From the definition it follows

$$\begin{aligned} P(A \cap B) &= P(A/B)P(B) \\ P(A \cap B) &= P(B/A)P(A) \end{aligned}. \qquad (1.9)$$

The Baye’s rule is then readily derived

Assuming that the event B is from the control parameter space and A from the measurement space, then the rule (1.7) is red as:

“*Conditional probability of the parameter taking values defined by the event B conditioned on the measurement taking values as defined by the event A is equal to product of conditional probability of the measurement taking values defined by the event A conditioned on the parameter taking values as defined by the event B and probability of the parameter taking the values as defined by the event B, normalized by probability of the measurement taking values as defined by the event A*

This relationship apparently allows to evaluate probability of the parameter (as defined by B) given the measurements (as defined by A) assuming that right hand side (r.h.s.) of (1.7) is known. The probabilities $P(A/B)$, $P(A)$ and $P(B)$ are hard to evaluate when based on the occurrence of events approach. It is far more convenient to assume probability distributions associated with the space to which the events A and B belong (i.e., the measurement and parameter spaces, respectively). As the distribution is determined by the probability density (1.6), the problem is then transformed to finding a relationship between the probability densities on the joint parameter and measurement spaces. To arrive at the relationship which relates the probability densities instead of the

probabilities of individual events we take the approach from Tarantola (2005) of defining the probability densities in the joint spaces of the parameters and measurements and their conjuction .

It is possible but not necessary to derive the desired relationship between the probability densities as generalization of the Bayes' rule (1.7). This approach is taken in the literature on estimation and stochasting filtering theory which addresses the inference of state of modeled time evolving systems from discrete stochastic measurements (Jazwinski, 1970; Sorenson, 1985). In the applications in the Geoscience problem examples of the use of equivalent to the Bayes' rule for probabilities is described in Cohn (1997), Rodgers (2000), Evensen (2006) and Lewis et. al. (2005). In the stochastic filtering theory literature the generalization of the Bayes' rule is derived by a limiting process in the joint space of the measurements and modeled state (Jazwinski, 1970). It is beyond the scope of this text to present the theoretical derivation and indebt analysis of the use of conditional instead of the joint posterior probability density functions. In the present chapter the approach from Tarantola (2005) is adapted for easy interpretation of the origin of probability density functions on the parameter and measurement spaces which apply within wide scope of the Geoscience problems where the parameters and models of many kinds are used to analyze and predict the state in conjunction with vast variety of measurements.

2.4.2 Conjuction of probability distributions

It is shown in section 2.1 that there are two sources of information about the natural system under study. These are the modeled and measured information. Let parameter space be denoted M, spanned by points $(m_1, m_2, ...)$. This space is transformed into a measurement space by a forward model

$$y = \phi(m) \qquad (1.10)$$

In the damped oscillator example ϕ is represented by equation (1.1). Let the measurement space as simulated by the model be denoted O. O is spanned by points $(y_1, y_2, ...)$. The joint space $\Omega = M \times \mathrm{O}$, which is characterized by a joint probability density $f(m, y)$, is the space of all possible information available about a natural system under study, given the model. The joint probability density on Ω provides complete description of the uncertainties and natural variability in the parameters and the result of these by the model simulations which is contained in the space O. The joint probability density $f(m, y)$ could also include effects of modeling errors. The modeling errors would result from the use of imperfect model. For example, the damped oscillator model (1.1) may be used to simulate damped oscillations which are in reality also driven by some unknown external harmonic force. When the force is not included in the equation, the model would be in error relative to the actual natural system and consequently relative to the measurements. It is not trivial task to design or assume the effect of modeling errors when specifying the joint probability density $f(m, y)$. This problem is illustrated in the exercises ?? .

The other information about the natural system is contained in the actual measurements which are independent of the model. Let this information be in space denoted C. There is a joint probability density on the joint space $\Theta = C \times M$, denoted $\rho(m,d)$. Notice that $\rho(m,y) \neq f(m,y)$. The union of measurement spaces O and C defines total measurement space which is denoted D. The joint probability densities $f(m,y)$ and $\rho(m,y)$ are both defined on D. New information about the system would be obtained when the two joint probability densities are combined by conjuction (Tarantola, 2005, chapter 1.5)

$$p(m,y) = \frac{1}{\gamma}\frac{\rho(m,y)\,f(m,y)}{\nu(m,y)} \qquad (1.11)$$

Where $\gamma = \int_{D\times M} \frac{\rho(m,y)\,f(m,y)}{\nu(m,y)}$ is constant and ν is so called homogenous probability density. $p(m,y)$ is ***a posteriori*** probability density on the joint space $D \times M$ resulting from the combined probability distributions. The knowledge of a posteriori probability density is the most complete available quantitative knowledge of information about the natural system under study. By this property, the expression (1.11) defines the general inverse modeling problem:

Evaluate $p(m,y)$ from knowledge of $\rho(m,y)$, $f(m,y)$ and $\nu(m,y)$.

$p(m,y)$ contains all available quantitative information of the system in the space $M \times D$ from which solution of the inverse modeling problem is to be derived. To arrive

at the resolution we need to introduce definitions of marginal and conditional probability densities and a priori information.

The marginal probability densities for any joint space with the associated joint distribution $g(a,b)$ in the space spanned by points $(a_1,a_2,.....,b_1,b_2,.....)$ are

$$g_A(a) = \int_B g(a,b)db$$
$$g_B(b) = \int_A g(a,b)da \qquad (1.12)$$

When the space $(a_1,a_2,.....)$ is independent of $(b_1,b_2,.....)$ then

$$g(a,b) = g_A(a)g_B(b) \qquad (1.13)$$

The conditional probability density is defined

$$g_{A/B}(a/b) = \frac{g(a,b)}{g_B(b)} \qquad (1.14)$$

The conditional probability density is interpreted as the probability density of points in the joint space for which $b = b(a)$. Using (1.14) the joint probability density for the information given the model is

$$f(m,y) = f(y/m)\nu(m) \qquad (1.15)$$

where the marginal probability density in the parameter space is assumed to be equal to the homogenous probability density of the parameters. The conditional probability density $f(y/m)$ is made of the results of forward model applied over a space of control parameters **without knowledge of the measurements.** In figure 1.2 the discrete examples of this probability density are shown for the damped oscillator model.

The probability density $\rho(m, y)$ results from the information in the joint space of the control parameters and measurements without knowledge of the model. It is natural to assume that these are independent. From (1.13)

$$\rho(m, y) = \rho_M(m)\rho_D(y) \tag{1.16}$$

The probability density $\rho_D(y)$ results exclusively from information about the uncertainties in the measurements. The probability density $\rho_M(m)$ in turn results from information of the uncertainties or variability in the control parameters which is independent of the measurements. This information is called ***a priori***.

Under the same assumption as in (1.16) the homogenous probability density in (1.11) is

$$\nu(m, y) = \nu_M(m)\nu_D(y) \tag{1.17}$$

Substituting (1.15-1.17) into (1.11) renders

$$p(m, y) = \frac{1}{\gamma}\frac{\rho_D(y)\rho_M(m) f(y/m)}{\nu_D(y)} \tag{1.18}$$

The solution of the general problem (1.18) is to compute the marginal probability density for the control parameter space. Using (1.18) in (1.12)

$$p_M(m) = \int_D \frac{1}{\gamma} \frac{\rho_D(y)\rho_M(m) f(y/m)}{\nu_D(y)} dy \qquad (1.19)$$

(Tarantola, 2005). $p_M(m)$ is interpreted as projection onto M. The probability densities on the r.h.s. of (1.19) are assumed known for the specific application. Geosciences and other physical sciences where parameters and models of many kinds are used to analyze and predict the state in conjunction with vast variety of measurements.

In the present definition of the joint space $M \times D$ with the associated joint probability density $p(m,d)$, the conditional probability density in for the parameter space could be derived from application of (1.14) assuming existence of $m = m(y)$. This assumption is somewhat difficult to interpret in the general case in which the parameter space is not the same as the modeled system state as in the stochastic filtering theory. When the assumption is valid it implies

$$p(m, y) = p(m/y)\, p_D(y) \qquad (1.20)$$

Combining (1.18) and (1.20)

$$p(m/y) = \frac{1}{\gamma} \frac{\rho_M(m)\rho_D(y) f(y/m)}{\nu_D(y) p_D(y)} \qquad (1.21)$$

This expression implies that the conditional probability density of parameters conditioned on the measurements is obtainable from the independent information about quantities in the space of measurements and parameters. The posterior probability

densities in (1.19) and (1.21) are apparently different, but in either case the required knowledge about the independent stochastic information in the parameter and measurements spaces is the same. Before addressing common choices in general and more specifically in the examples in chapters 4 and 5 it is instructive to consider what type of information may be most useful or interesting to derive from the knowledge of posterior probability density function.

2.4.3 Estimation criteria

In the practice with the Geoscience problems the parameter space is often large multidimensional space. In this situation it is unfeasible to either evaluate or visualize (1.19) or (1.22). Instead, characteristics of the posterior probability density function are used to define single ***best estimate*** or ***central estimate*** of the parameters (Cohn, 1997, Jazwinski 1970; Tarantola, 2005). The commonly used central ***estimation criteria*** are

a) ***Maximum likelihood,*** define by a discrete region or continuous point with maximum probability associated with the posterior probability density function. The likelihood function is

$$L(m) = \int_D \frac{\rho_D(y) f(y/m)}{\nu_D(y)} dy$$

b) ***Minimum variance,*** defined by the men of the posterior probability distribution

$\langle m \rangle = \int_M m p_M(m) dm$ or conditional mean $\langle m \rangle = \int_M m p(m/y) dm$

c) ***Minimum absolute distance,*** defined by the median of the posterior density distribution

The choice of criterion would depend on the purpose of estimation and characteristics of the specific problem.

2.4.4 Conjuction of Gaussian distributions

It is common to assume that probability density functions associated with model and measurement spaces are Gaussian or Normal. The Gaussian distribution is characterized with only two statistical parameters: mean $\langle x \rangle$ and covariance C

$$p(x) = \frac{1}{(2\pi)\det^{\frac{1}{2}} C} \exp\left(-\frac{1}{2}(x - \langle x \rangle)^T C^{-1}(x - \langle x \rangle) \right) \qquad (1.23)$$

In the model space (1.23) is

$$f(y/m) = \frac{1}{(2\pi)\det|C_s|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(y - \phi(m))^T C_s^{-1}(y - \phi(m)) \right) \qquad (1.24)$$

while in the measurement space

$$\rho_D(y) = \frac{1}{(2\pi)\det|C_d|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(y - y_{meas})^T C_d^{-1}(y - y_{meas})) \right) \qquad (1.25)$$

Using (1.24) and (1.25) in (1.19)

$$p_M(m) = k\rho_M(m)\exp\left(-\frac{1}{2}(\phi(m) - y_{meas})^T C_D^{-1}(\phi(m) - y_{meas})\right) \tag{1.26}$$

Where d_{meas} denotes actual measurements, k is cumulative constant and

$$C_D = C_d + C_S \tag{1.27}$$

(1.26) indicates that the convolving of the two Gaussian probability distributions in the measurement space is also Gaussian with the summed up uncertainties from the independent modeled and measured information, represented by the covariance C_D

Problem 1: Derive 1.27 (Appendix)

When it is further assumed that the a priori probability density function in the parameter space is Gaussian

$$\rho_M(m) = \frac{1}{(2\pi)\det|C_m|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(m - m_{prior})^T C_m^{-1}(m - m_{prior}))\right) \tag{1.28}$$

then

$$\begin{aligned} p_M(m) &= const\exp(-S(m)) \\ S(m) &= \frac{1}{2}\left[(\phi(m) - y_{meas})^T C_D^{-1}(\phi(m) - y_{meas}) + (m - m_{prior})^T C_m^{-1}(m - m_{prior})\right] \end{aligned} \tag{1.29}$$

$S(m)$ is apparently the weighted sum of squares. When the model is linear $\phi(m) \equiv Fm$, then $p_M(m)$ in (1.29) becomes Gaussian with the mean and covariance, respectively

$$\begin{aligned} \langle m\rangle &= m_{prior} + C_M F^T\left(FC_M F^T + C_D\right)^{-1}(d_{meas} - Fm_{prior}) \\ C_S &= (F^T C_D^{-1} F + C_M^{-1})^{-1} \end{aligned} \tag{1.30}$$

Problem 2: Derive 1.30 (Appendix)

In the section on Kalman Filter technique (3.2) it is shown that solution (1.30) is also derived for the data assimilation problem by the stochasting filtering theory which addresses the inference of state of modeled time evolving systems from discrete stochastic measurements (Jazwinski, 1970). This theory is applicable in the Atmospheric sciences and Oceanography when the interest is to produce quantification of the atmospheric or oceanic state in geographical discretized space and over time (Cohn, 1997; Kalnay 2000).

Application of the maximum likelihood criterion for the central estimate by (1.29) implies minimization of $S(m)$. The minimization of $S(m)$ is commonly referred to as "least square problem" which is treated in the chapter on variational techniques (3.3).

CHAPTER 3 - DRAFT

Inverse Modeling and Data Assimilation in Geosciences: A Practial Guide

Author : Dr. Tomislava Vukicevic (*tomislava.vukicevic@colorado.edu*)

Affiliation : Associate Professor, Department of Atmospheric and Oceanic Sciences (ATOC), University of Colorado, Boulder, Colorado, USA

Date of this document: December 2006

---

3. Inverse modeling and data assimilation techniques

3.1. Monte Carlo

(Missing brief intro: bit of history, mention of kind of problems which are solved with this class of techniques)

It is easy to understand that when a space of stochastic quantity is sampled randomly **many times** it is possible to derive a distribution on that space from the sample. This is the desired result in the inverse problem, if obtainable. The random sampling is called ***Monte Carlo*** sampling (Mannon, 1999). The difficulty is that for multidimensional spaces such as the space of parameters which is transformed into the simulated measurements in the inverse problems in the Geosciences, there are large regions of insignificant resultant probability, implying need to have very large samples. To reduce the number of samples, it is desirable to tend to sample regions which result in

significant probability. Several sampling techniques which posses this property are briefly described in this section.

Even when made "efficient" the Monte Carlo techniques are practical only for problems with relatively small number of parameters (order of tens or less) and relatively fast models. In the exercises (chapter 5) the reader could experience performance of an efficient Monte Carlo inverse technique, called Markov Chain Monte Carlo (MCMC) by Metropolis and Ulam (1949) with models of varying degree of computational cost and complexity. Compared to other techniques included in this practicum the MCMC is by far the most costly.

3.1.1. Metropolis

In the inverse problem the interest is to sample $p_M(m)$ using (1.19). This implies sampling of the conjuction of $\rho_D(d)$ and $f(d/m)$ from random independent samples with probability distribution $\rho_M(m)$. One of most efficient techniques to do this is called ***Metropolis*** (Metropolis and Ulam (1949). First it is assumed that each step in sampling is dependent only on the previous step. This is called ***Markov Chain*** (reference ?). Second, the sampling is random at each step which is characteristic of the Monte Carlo sampling but the move from one step to the next is controlled in the following way

- If $L(m_j) \geq L(m_i)$ then accept the transition from m_i to m_j
- If $L(m_j) < L(m_i)$ then decide to randomly move to m_j or to stay at m_i with the following probability of accepting the move

$$P_{i\to j} = \frac{L(m_j)}{L(m_i)} \qquad (1.31)$$

where $L(m) = \int_D \frac{\rho_D(y) f(y/m)}{\nu_D(y)} dy$ is the likelihood function. Obviously, there has to be an initial estimate of $f(y/m)$ to be able to evaluate the likelihood function for the test at (m_i, m_j). $\rho_D(y)$ is typically not produced by sampling but is assumed to be known density derived from knowledge about the specific measurement uncertainties. The initial estimate of $f(y/m)$ could result from : a) an independent random sampling, typically referred to as "burn in period" or b) approximation by a known density distribution function such as the Gaussian.

In the Gaussian case the likelihood function from (1.26) applies but the associated covariance is not known exactly. When it is further assumed that the measurements are independent random quantities the requirement to use an approximate covariance results in simpler requirement to specify an initial approximate variance for each measured quantity. The approximated likelihood function is then

$$\tilde{L}(m) = const \exp\left[(\phi(m) - y_{meas})^T \Lambda^{-1} (\phi(m) - y_{meas})\right] \qquad (1.32)$$

where Λ is diagonal matrix of the combined measurement and model uncertainties as in (1.27). During the random sampling using the criteria of the Metropolis technique the approximated likelihood function could be refined by introducing new estimates of the variance.

3.1.2. Simulated Annealing

When the goal is to produce just the maximum likelihood estimate by random sampling, the technique analogous to the physical process of annealing could be used. The physical annealing consists of first heating than slow cooling of solids to ambient temperature to eliminate stress in the material. In the numerical technique labeled the "simulated annealing" an energy function is defined

$$E(m) = -T\ln(L(m)) \qquad (1.33)$$

where T is equivalent of temperature. The energy is always positive. The posterior probability density function

$$p_M(m) = const\rho_M(m)e^{-\frac{E(m)}{T}}$$

is at maximum when the energy function is at minimum. The technique consists of slow change of "temperature" toward zero to render the energy minimum. The energy function for the probability density distribution in (1.27) which results from the conjuction of Gaussian distributions is the misfit function in the measurement space multiplied by a constant.

$$E(m,\mathrm{T}) = \mathrm{T}k\left((\phi(m) - y_{meas})^T C_D^{-1}(\phi(m) - y_{meas})\right) \qquad (1.34)$$

The assumption of independent measurements would render the covariance C_D diagonal as in (1.32). The advantage of (1.32) or (1.34) is that the conjuction is explicitly and easily evaluated for any m. To test applicability of (1.32) and (1.34) or to estimate the approximate variance it is desirable, if feasible, to "roughly" evaluate

$f(y/m)$ or the conjuction $\rho_D(y)f(y/m)$ by a simpler Monte Carlo sampling technique. For example, this could be performed by the Gibbs technique (Geman and Geman, 1984).

3.2.Kalman Filter

The Kalman Filter class of techniques have been developed for solving the problems of system control by estimation of time evolving state of the system from erroneous measurements. In this class of problems the state of the system at one time controls the state at subsequent time (Kalman, 1960; Jazwinski, 1970). The Kalman Filter techniques have been introduced in 1990's in the Geoscience disciplines to address similar problem. An excellent introduction to the application of Kalman Filter techniques in the Atmospheric sciences is given in the article by Cohn (1997).

Central to the Kalman Filter class of techniques are the use of posterior conditional probability density and assumption that the prior, modeled and actual measurement probability densities are Gaussian. To connect to the general inverse problem theory as presented in chapter 2 following Tarantola (2005) recall that the posterior conditional probability density resulting from conjuction of information on the space $M \times D$ is expressed by (1.21), from which it follows

$$p(m/y) = \frac{p(m,y)}{\int_D p(m,y)dy} = \frac{p(y/m)p_M(m)}{\int_D p(m,y)dy} \qquad (1.35)$$

The relationship (1.35) is then used as the solution of the general inverse problem (1.11). It is already discussed in chapter 2 that the use of either the marginal or

conditional posterior probability density does not change the inverse problem. The problem is to find the probability density function which combines prior knowledge about the stochastic control parameters employed by the model with the stochastic measurements.

The assumption of Gaussian probability densities is less general. It is reasonable to pose the question: When is the use of Gaussian probability density valid? The ***Central Limit Theorem*** on the properties of cumulative stochastic quantities partially helps in answering the question. It states

Central Limit Theorem: *Cumulative distribution of any set of independent variables with any distribution having a finite mean and variance tends to normal distribution*

This theorem is readily interpreted in the measurement space as a large number of different measurements of the same physical quantity would tend to produce normally distributed cumulative estimate of that quantity. The problem may occur with measured quantities which are positive semi-definite ($y \geq 0$) with large probability near or at zero. A transformation of variable

$$q = \ln(y) \qquad (1.37)$$

could help solve the problem if the associated probability density function $\rho(q)$ is exactly or approximately Gaussian. The validity of Gaussian probability density assumption in the parameter and model spaces is hard to justify for general case and must be addressed for the each specific problem. Transformations similar to (1.37) may be used.

Assuming the validity of Gaussian probability density functions and the use of conditional probability density as the solution of inverse problem the theory of discrete Kalman Filter for linear class of models is derived in the following way. Let a discrete dynamical or state evolution model be

$$w_k^t = F_{k-1} w_{k-1}^t + G_{k-1} \varepsilon_{k-1}^t \qquad (1.38)$$

where w_k^t is a vector of numerical solution of a set of differential equations in discrete time and space. Subscript k denotes discrete time point, while superscripts t and o refer to ***truth*** and ***observations***, respectively. The symbol w is used to denote the control parameters instead of m to distinguish the specific type of the control parameter in the state estimation problem. The equation (1.38) expresses transformation of the "true" state from time step $(k-1)$ to k by the discrete linear operator F_{k-1} and the associated error relative to the true transformation . This error is assumed to be a linear stochastic forcing $G_{k-1}\varepsilon_{k-1}^t$. In what follows the exact knowledge of deterministic error relative to the truth is not required but the knowledge of error statistics is.

The state evolution model typically does not integrate into the observation space. This condition requires that the state is transformed or mapped as in (1.10). In the state estimation problem the transformation is written

$$y_k^o = H_k w_k^t + \varepsilon_k^o \qquad (1.39)$$

This transformation uses a discrete linear operator H from the true state to the observations. The linear operator is assumed because the original Kalman Filter was derived for the linear transformations. The linear term ε_k^o in (1.39) indicates that there

are errors in the observations or, in terms of the formulation in chapter 2, that the observations are stochastic quantities. This error term could also include the errors which are associated with uncertainties in the transformation operator.

Substitution of (1.39) into (1.38) would render one model for simulation of transformation from the state at $k-1$ time to measurements at k. The separation of the model in two parts is motivated with goal to produce estimates of the state sequentially at discrete time points as new measurements become available after the prediction is made by the evolution model (1.38) using an estimate of the true state at previous times. The evolution between $(k-1)$ and (k) from the estimate is written

$$w_k^f = F_{k-1} w_{k-1}^e + G_{k-1} \varepsilon_{k-1}^e \qquad (1.40)$$

The predicted state is then transformed into the measurement space

$$y_k^o = H_k w_k^f + \varepsilon_k^f \qquad (1.41)$$

The inverse problem is to find an estimate of the true state w_k^t from information about stochastic quantities w_k^f, y_k^o, ε_k^t and ε_k^o. The conditional probability (1.36) is used for the solution. The prior probability density $\rho_M(m)$ is the probability density of the predicted state in (1.40). The probability density of measurements given the model is the probability density of y_k^o in (1.41). The marginal probability density of the measurements is $p_D(y_k^0)$. The posterior conditional probability density is then written

$$p(w_k^t / y_k^o) = const \frac{p(w_k^t / y_{k-1}^o) p(y_k^o / w_k^t)}{p(y_k^o / y_{k-1}^o)} \qquad (1.42)$$

The conditional probability density notation for the prior and measurement terms on the right hand side is used to indicate that the knowledge about them is conditioned on the observations which were used prior to making the new prediction. This property implies that the solution (1.42) is recursive. Assuming that all probability densities on the right hand side in (1.42) are Gaussian, the solution requires that the associated mean and covarinaces are specified.

From (1.40) the mean and covariance of $p(w_k^t / y_{k-1}^o)$ are derived, respectively

$$\begin{aligned}\left\langle w_k^t / y_{k-1}^o \right\rangle &= F_{k-1}\left\langle w_{k-1}^t / y_{k-1}^o \right\rangle + G_{k-1}\left\langle \varepsilon_{k-1}^t / y_{k-1}^o \right\rangle = F_{k-1} w_{k-1}^e \\ C_k &= \left\langle \left[F_{k-1}(w_{k-1}^t - w_{k-1}^e) + G_{k-1}\varepsilon_{k-1}^t\right]\left[F_{k-1}(w_{k-1}^t - w_{k-1}^e) + G_{k-1}\varepsilon_{k-1}^t\right]^{\mathrm{T}} \right\rangle \\ &= F_{k-1} C_{k-1} F_{k-1}^{\mathrm{T}} + G_{k-1} Q_{k-1} G_{k-1}^{\mathrm{T}}\end{aligned} \quad (1.43)$$

Here it is assumed that the mean of the state evolution model error is identically zero (i.e., the error is the "white noise") and that the mean of the probability density at $(k-1)$ is the central posterior estimate at that time, which is then propagated forward in time. The mean and covariance of $p(y_k^o / w_k^t)$ and $p(y_k^o / y_{k-1}^o)$ are derived using (1.41) and (1.43) along with assumption that the model error is white and that the estimate and model errors are mutually uncorrelated.

$$\begin{aligned}\left\langle y_k^o / w_k^t \right\rangle &= H_k w_k^t \\ R_k &= \left\langle \left(y_k^o - H w_k^t\right)\left(y_k^o - H w_k^t\right)^{\mathrm{T}} \right\rangle\end{aligned} \quad (1.44)$$

$$\begin{aligned}\left\langle y_k^o / y_{k-1}^o \right\rangle &= \left\langle H_k w_k^t + \varepsilon_k^o \right\rangle = H_k w_k^f \\ \left\langle \left(y_k^o - H_k w_k^f\right)\left(y_k^o - w_k^f\right)^{\mathrm{T}} \right\rangle &= H_k C_k H_k^{\mathrm{T}} + R_k\end{aligned} \quad (1.45)$$

Using (1.43)-(1.45) in the definition of Gaussian probability density function and substituting in (1.42) results in the expression for the desired posterior conditional probability density

$$p(w_k^t / y_k^o) = const \exp(-\frac{1}{2} J) \tag{1.46}$$

where

$$\begin{aligned} J &= (y^o - Hw^f)^{\mathrm{T}} R^{-1} (y^o - Hw^f) + (w^t - w^f)^{\mathrm{T}} C^{-1} (w^t - w^f) \\ &\quad - (y^o - Hw^f)^{\mathrm{T}} (HP_f H^{\mathrm{T}} + R)^{-1} (y^o - w^f) \end{aligned} \tag{1.47}$$

It can be shown (for example in Cohn, 1997) that

$$J = \left[w_k^t - \left(w_k^f + K_k (y_k^o - H_k w_k^f)\right)\right]^{\mathrm{T}} (C_k^e)^{-1} \left[w_k^t - \left(w_k^f + K_k (y_k^o - H_k w_k^f)\right)\right] \tag{1.48}$$

$$C^e = \left(H^{\mathrm{T}} R^{-1} H + C^{-1}\right)^{-1} \tag{1.49}$$

$$K = CH^{\mathrm{T}} \left(HCH^{\mathrm{T}} + R\right)^{-1} \tag{1.50}$$

Substitution of (1.48)-(1.50) shows that the posterior conditional probability density function is Gaussian with mean and covariance, respectively

$$\begin{aligned} w_k^e &= w_k^f + C_k H^{\mathrm{T}} \left(H_k C_k H_k^{\mathrm{T}} + R\right)^{-1} \left(y_k^o - H_k w_k^f\right) \\ C_k^e &= \left(H_k^{\mathrm{T}} R^{-1} H + C_k^{-1}\right)^{-1} \end{aligned} \tag{1.51}$$

The matrix K in (1.50) is called ***Kalman gain*** matrix because it is applied to the difference between actual and simulated measurements to correct the prior in (1.51). The mean and covariance in (1.51) could be obtained directly from the conjuction of Gaussian probability densities in section 2.4.4 as application of (1.30) to the problem in this section where the prior is produced sequentially from prediction using the posterior solution at previous time instance.

The most celebrated property of the Kalman Filter is that the entire probability density function is predicted to produce the new prior because both moments of the Gaussian probability density function are advanced in time by (1.43). The Kalman Filter is however difficult to apply exactly even for exactly linear models when the number of elements N in the state vector w_k^f is large as it requires prediction of the covariance matrix of the size N^2. The filter could be applied exactly to smaller size problems (order of 100) with efficiency that may exceed the performance of a Monte Carlo type algorithm. Feasibility of the Kalman Filter theory is extended by introduction of the ensemble approach in section 3.4.

3.3. Variational techniques

It is shown in section 2 that posterior probability density function (pdf) is expressed

$$p_M(m) = const \exp\left(-S(m)\right)$$

$$S(m) = \frac{1}{2}\left[\left(\phi(m) - y_{meas}\right)^T C_D^{-1}\left(\phi(m) - y_{meas}\right) + \left(m - m_{prior}\right)^T C_m^{-1}\left(m - m_{prior}\right)\right]$$

(1.52)

under condition that the prior, modeled and actual measurement stochastic quantities have the Gaussian pdfs. This condition is used in derivation of the Linear Kalman Filter technique in section 4 together with assumption that the models which represent evolution of a dynamical system and the mapping or transformation into the measurement space are both linear. These conditions together render solution of the inverse problem

in the model state space ($m = w_k^e$) with Gaussian posterior probability density function. The mean of posterior distribution in this case is the minimum of the cost function $S(m)$. The Linear Kalman Filter technique is not, however, widely used in the Geosciences because it requires numerical evaluation and inversion of very large matrices and because it is derived for strictly linear model. Instead of evaluating the entire posterior Gaussian pdf, the inverse problem could be reduced to solving for a central estimate such as the mean or maximum. In section 2.3 it is shown that the mean is central estimate under minimum variance criterion for any distribution. In the Gaussian posterior pdf case the mean and maximum are the same. This property implies that the mean also satisfies maximum likelihood criterion.

If the condition of linear model is eliminated, which eliminates validity of the Gaussian posterior pdf, but the prior, modeled and actual measurement pdfs remain Gaussian the minimum of cost function would still satisfy the maximum likelihood criterion. This property suggests the use of cost function minimum to obtain the maximum likelihood central estimate for unknown posterior distribution. The inverse problem of this kind is more general than the problem addressed in the Linear Kalman Filter but less general in solution as it does not attempt to evaluate the entire posterior pdf.

The minimization of a quadratic misfit function such as the cost function in 1.51 defines an entire class of problems in the estimation theory which is commonly referred to as Least Square problem (Crassidis and Junkins, 2004; Lewis et al, 2006; Tarantola, 2005). The least square problem was first introduced by Gauss (1809) to determine

planetary orbits from an orbit model and telescope measurements of the line of sight angles. In this original work the sum of squares of un-weighted differences between model and observations is minimized. This kind of least square problem is commonly referred to in modern literature as unconstrained least square problem without weights and prior. In the cost function (1.51) the absence of weights is equivalent to assuming that $C_D = I$ (I is identity matrix), while no-prior implies $C_M^{-1} = 0$. The unconstrained property implies that the minimization is performed directly on the cost function without additional functional relationship on the parameter space. The least square problem defined by (1.51) is thus referred to as constrained, weighted with prior. The constraint is provided by a model such as prediction model (1.38) in the previous section.

The minimization of cost function could be desirable problem to solve even for the linear models if the inversion of large matrices could be avoided. The least square problems (linear and nonlinear) could be solved relatively efficiently by Variational technique (Kalnay, 2004, Lewis et al, 2005). The solution is obtained by use of the variational calculus which involves evaluation of directional gradients of the cost function (Crassides and Junkins, 2005; Tarantola, 2005; Bryson and Ho, 1975). There are numerous minimization techniques described in the literature on optimal estimation and control theory which make use of the cost function directional gradients. In essence they use the following well known properties of functions

Necessary and sufficient condition for a minimum of a function $f(x)$ which is at least twice differentiable are, respectively

$$\frac{\partial f}{\partial x} = 0 \text{ and } \frac{\partial^2 f}{\partial x^2} > 0$$

where x is vector of independent variables (equivalent of parameter space). Second order truncated Taylor series expansion for $f(x)$ at an arbitrary point x^* reads

$$f(x) = f(x^*) + \left[\frac{\partial f}{\partial x}\right]_{x=x^*}(x - x^*) + \frac{1}{2}\left[\frac{\partial^2 f}{\partial x^2}\right]_{x=x^*}(x - x^*)^2 \qquad (1.53)$$

Differentiation with respect to x and setting $x = x_{\min}$, together with the necessary condition yields

$$x_{\min} = x^* - \left[\frac{\partial^2 f}{\partial x^2}\right]^{-1}_{x=x^*}\left[\frac{\partial f(x)}{\partial x}\right]_{x=x^*} \qquad (1.54)$$

This relationship shows that a minimum of function is directly proportional to negative of the first directional derivative at an arbitrary point x^* in the neighborhood of the minimum. The use of second order Taylor expansion approximation is made possible by assumption that this point is not far from the minimum. For the exactly quadratic function, the second order truncation is exact. The cost function in 1.51 is exactly quadratic if the model is linear and approximately quadratic in the neighborhood of the minimum if the model is nonlinear. Consequently, the relationship (1.54) may be used to find the minimum if these conditions are satisfied. It is obvious that unless the function is exactly quadratic, the minimum obtained in this way may not be global.

The nature of minimum should be examined for each application specifically if possible by inspection of extent of the neighborhood around the computed minimum within which $S(x_j) > S(x_{\min})$. The neighborhood should be spanning the range of parameter values with large cumulative probability. This condition is difficult to inspect

for large multi dimensional problems. A natural test of improvement brought about by results of the inverse problem with the prognostic type model may be to verify the prediction against new measurements, assuming that the modeling errors are small within forecast temporal range. If the prediction is better with than without the solution of the inverse problem, the inverse technique has skill, implying that the assumptions used in the technique are at least approximately valid.

3.3.1 Variational solution of constrained minimization problem with prior

In chapter 1 it is discussed that many models used in the Geosciences simulate time evolution of the natural system state. Assume that the model of interest is written as system of ODE-s

$$\frac{d\chi}{d\tau} = M\left(\chi,\alpha\right) + G(\varepsilon,\tau) \qquad (1.55)$$

The models which do not simulate time evolution could be written in the same symbolic form but with $\frac{d\chi}{d\tau} = 0$, implying steady state. The model (1.55) is equivalent to the prognostic model (1.38) in section 3.2, but the time derivation is expressed in the continuous form. The model and model error operators $M(\chi,\alpha)$ and $G(\varepsilon,\tau)$, respectively, are assumed in general nonlinear. The system state vector of physical quantities $\chi(\tau)$ is in discretized space. The solution of (1.55) is subject to initial and boundary conditions, respectively

$$\chi_{\tau-\Delta\tau} = \chi_{prior} \quad , \quad \chi(\Omega) = \chi_b(\tau), \qquad (1.56)$$

where Ω denotes boundary of the spatial domain. The solution is also dependent on vectors of physical parameters $\alpha(\tau)$ and model error $\varepsilon(\tau)$. In the inverse problem the measurements are contained in a vector of measured quantities in discrete spatial points at discrete times $\tau_k \in (\tau - \Delta\tau, \tau)$, as in section 4. The transformation from the system state space into measurement space is as in section 4

$$y(\tau_k) = h(\chi_k) + \varepsilon_k^o \tag{1.57}$$

Unlike in the linear Kalman Filter, the transformation function h is assumed general nonlinear. The cost function (1.51) for the system (1.55-1.57) reads

$$\begin{aligned} S = \frac{1}{2}\int_{\tau-\Delta\tau}^{\tau} \Big[&\left(\alpha - \alpha_{prior}\right)^{\mathrm{T}} C_\alpha^{-1} \left(\alpha - \alpha_{prior}\right) + \left(\chi - \chi_b\right)^{\mathrm{T}} C_b^{-1} \left(\chi - \chi_b\right) + \varepsilon^{\mathrm{T}} Q \varepsilon \Big] \delta(\tau - \tau_k) d\tau \\ &+ \frac{1}{2}\left(\chi_{\tau-\Delta\tau} - \chi_{prior}\right)^{\mathrm{T}} C_f^{-1} \left(\chi_{\tau-\Delta\tau} - \chi_{prior}\right) \\ &+ \frac{1}{2}\int_{\tau-\Delta\tau}^{\tau} \Big\{ \big[h(\chi) - y_k\big] C_D^{-1} \big[h(\chi) - y_k\big] \delta(\tau - \tau_k) \Big\} d\tau \end{aligned}$$

(1.58)

$\delta(\tau - \tau_k)$ is Kroneker delta function in time. The prior is represented with four terms instead of one used in section 4, because the prognostic model solution depends on four types of control parameters. Each of these parameters could be varied to render the cost function minimum. The problem of finding the minimum of (1.58) under constraint given by (1.55) can be compactly written

$$\begin{aligned} S &= S(x) \\ \psi(x) &= 0 \end{aligned} \tag{1.59}$$

Where x is $m \times 1$ vector of all control parameters . The necessary condition at the minimum is $\frac{\partial S}{\partial x} = 0$. The condition for differential variation in the neighborhood of the minimum (Crassidis and Junkins, 2004) is then

$$\delta S = \frac{\partial S}{\partial x} \delta x = 0 \qquad (1.60a)$$

$$\delta \psi = \frac{\partial \psi}{\partial x} \delta x = 0 \qquad (1.60b)$$

The minimum solution is then obtained by elimination of differential variations in each component of x from (1.60b) and substitution into (1.60a). This could be very difficult to solve as the functional relationship in the constraint could be very complex and even not known explicitly.

Lagrange derived transformation of the constrained minimization problem (1.60) into unconstrained by linearly combining equations in (1.59) to define new augmented functional

$$F = S + \lambda^{\mathrm{T}} \psi \qquad (1.61)$$

λ is Lagrange multiplier. The necessary conditions at the minimum of F read

$$\frac{\partial F}{\partial x} \delta x = \left\{ \frac{\partial S}{\partial x} + \left[\frac{\partial \psi}{\partial x} \right]^T \lambda \right\} \delta x = 0 \Rightarrow \frac{\partial S}{\partial x} + \left[\frac{\partial \psi}{\partial x} \right]^T \lambda = 0 \qquad (1.62)$$

$$\frac{\partial F}{\partial \lambda} \delta \lambda = \psi(x) = 0 \qquad (1.63)$$

(1.62) is system of equations for unknown λ, while the second condition recovers the original constraint. Solving the dual system (1.62-1.63) is an elegant way of automatic differential elimination.

A simple example from (Crassidis and Junkins, 2004) illustrates this property

Example #?

Find minimum of

$$S = 6 - \frac{y}{2} - \frac{z}{3}$$

under constraint $\psi(y,z) = 9(y-4)^2 + 4(z-5)^2 - 36 = 0$

The differential elimination by solving the dual system (1.63-1.64) is not applied explicitly in practice with large multidimensional problems but the solution of the system (1.61) is used. In what follows it is shown that for the large multidimensional control parameter space the vector of Largange multipliers is obtained from the solution of 1.61 and that this solution is directly proportional to the vector of directional gradients of the cost function with respect to each control variable. This property allows numerical evaluation of the directional gradient vector which could be then used in the relationship (1.54) or similar to compute the minimum.

To derive (1.61) explicitly for the system (1.55-1.58) define first the augmented functional

$$F = S + \int_{\tau-\Delta\tau}^{\tau} \lambda(\tau)\left(\frac{d\chi}{d\tau} - M(\chi,\alpha) - G(\varepsilon,\tau)\right)d\tau \tag{1.64}$$

The Lagrange multiplier is a vector of the same size as χ. The variation of augmented functional by the variation of what controls χ is as follows

$$\begin{aligned}
&\delta F = \delta F_1 + \delta F_2 \\
&\delta F_1 = \delta S \\
&\delta F_2 = \delta\left[\int_0^{\tau} \lambda(\tau)\left(\frac{d\chi}{d\tau} - M(\chi,\alpha) + G(\varepsilon,\tau)\right)d\tau\right]
\end{aligned}$$

$$\begin{aligned}
\delta F_1 = & \underbrace{\int_{\tau-\Delta\tau}^{\tau} \delta\alpha^{\mathrm{T}}\left\{C_\alpha^{-1}\left[\alpha-\alpha_{prior}\right] + \left[\frac{\partial\chi}{\partial\alpha}\right]^T\left[\frac{\partial h}{\partial\chi}\right]^T C_D^{-1}\left[h(\chi)-y\right]\delta(\tau-\tau_k)\right\}d\tau}_{A1} \\
& \underbrace{+\int_{\tau-\Delta\tau}^{\tau} \delta\varepsilon^{\mathrm{T}}\left\{Q^{-1}\varepsilon + \left[\frac{\partial\chi}{\partial\varepsilon}\right]^{\mathrm{T}}\left[\frac{\partial h}{\partial\chi}\right]^{\mathrm{T}} C_D^{-1}\left[h(\chi)-y\right]\delta(\tau-\tau_k)\right\}d\tau}_{A2} \\
& \underbrace{+\int_{\tau-\Delta\tau}^{\tau} \delta\chi_\Omega^{\mathrm{T}}\left\{C_b^{-1}(\chi_\Omega-\chi_b) + \left[\frac{\partial\chi}{\partial\chi_\Omega}\right]^{\mathrm{T}}\left[\frac{\partial h}{\partial\chi}\right]^{\mathrm{T}} C_D^{-1}\left[h(\chi)-y\right]\delta(\tau-\tau_k)\right\}_\Omega d\tau}_{A3} \\
& \underbrace{+\delta\chi_{\tau-\Delta\tau}^{\mathrm{T}} C_f^{-1}(\chi_{\tau-\Delta\tau}-\chi_{prior}) + \delta\chi_{\tau-\Delta\tau}^{\mathrm{T}}\int_{\tau-\Delta\tau}^{\tau}\left[\frac{\partial\chi}{\partial\chi_{\tau-\Delta\tau}}\right]^{\mathrm{T}}\left[\frac{\partial h}{\partial\chi}\right]^{\mathrm{T}} C_D^{-1}\left[h(\chi)-y\right]\delta(\tau-\tau_k)d\tau}_{A4}
\end{aligned} \tag{1.65}$$

$$\delta F_2 = \int_{\tau-\Delta\tau}^{\tau} \lambda\left[\frac{d\delta\chi}{d\tau} - \left[\frac{\partial M}{\partial \chi}\right]\delta\chi - \left[\frac{\partial G}{\partial \varepsilon}\right]\delta\varepsilon\right]d\tau$$

$$= \int_{\tau-\Delta\tau}^{\tau} \left\{\frac{d(\lambda\delta\chi)}{d\tau} - \delta\chi^{\mathrm{T}}\frac{d\lambda}{d\tau} - \delta\alpha^{T}\left[\frac{\partial M}{\partial \alpha}\right]\lambda - \delta\alpha^{T}\left[\frac{\partial M}{\partial \alpha}\right]\lambda - \delta\chi^{\mathrm{T}}\left[\frac{\partial M}{\partial \chi}\right]^{\mathrm{T}}\lambda - \delta\varepsilon^{\mathrm{T}}\left[\frac{\partial G}{\partial \varepsilon}\right]^{\mathrm{T}}\lambda\right\}$$

$$= \underbrace{[\lambda\delta\chi]_{\tau-\Delta\tau}^{\tau}}_{B1} + \underbrace{\int_{\tau-\Delta\tau}^{\tau} \delta\chi^{\mathrm{T}}\left(\frac{d\lambda}{d(-\tau)} - \left[\frac{\partial M}{\partial \chi}\right]^{\mathrm{T}}\lambda\right)d\tau}_{B2} - \underbrace{\delta\varepsilon^{\mathrm{T}}\left[\frac{\partial G}{\partial \varepsilon}\right]^{\mathrm{T}}\lambda d\tau}_{B3}$$

(1.66)

$$\delta\chi = \frac{\partial\chi}{\partial\alpha}\delta\alpha + \frac{\partial\chi}{\partial\chi_{\Omega}}\delta\chi_{\Omega} + \frac{\partial\chi}{\partial\chi_{\tau-\Delta\tau}}\delta\chi_{\tau-\Delta\tau} + \frac{\partial\chi}{\partial\varepsilon}\delta\varepsilon$$

(1.67)

Substituting (1.67) into (1.66) results in four terms within B2 in (1.66), one for each control parameter variation. When combined with A1-A4 in (1.65), it is easy to observe that there is common factor among contributions from each control parameter

$$I = \left(\frac{d\lambda}{d(-\tau)}\right) - \left[\frac{\partial M}{\partial \chi}\right]^{\mathrm{T}}\lambda + \left[\frac{\partial h}{\partial \chi}\right]^{T} C_D^{-1}[h(\chi) - y_k]\delta(\tau - \tau_k)$$

The variation of augmented functional is now written

$$
\begin{aligned}
\delta F = & \int \left\{ \delta\alpha^{\mathrm{T}} \left(C_{\alpha}^{-1}(\alpha - \alpha_{prior}) - \left[\frac{\partial M}{\partial \alpha} \right]^{\mathrm{T}} \lambda + I \right) \right\} d\tau \\
& + \int \delta\chi_{\Omega}^{\mathrm{T}} \left(C_{b}^{-1}(\chi_{\Omega} - \chi_{b}) - \left[\frac{\partial M}{\partial \chi_{\Omega}} \right]^{\mathrm{T}} \lambda + I \right) d\tau \\
& + \int \delta\varepsilon^{\mathrm{T}} \left(Q^{-1}\varepsilon - \left[\frac{\partial G}{\partial \varepsilon} \right]^{\mathrm{T}} \lambda + I \right) d\tau \\
& + \delta\chi_{\tau-\Delta\tau}^{\mathrm{T}} \left(\lambda(\tau - \Delta\tau) + C_{f}^{-1}(\chi_{\tau-\Delta} - \chi_{prior} \right) \\
& + \delta\chi_{\tau}^{\mathrm{T}} \lambda(\tau_{end})
\end{aligned}
\tag{1.68}
$$

At the minimum $\delta F = 0$, leading to a system of ODEs for λ as in (1.62). Given than λ is arbitrary, additional conditions may be set for λ without loss of generality. The condition $I = 0$ for all τ results in new system of ODEs for λ, called adjoint system.

$$
\frac{d\lambda}{d\tau^{*}} = \left[\frac{\partial M}{\partial \chi} \right]^{\mathrm{T}} \lambda - \left[\frac{\partial h}{\partial \chi} \right]^{T} C_{D}^{-1} \left[h(\chi) - y_{k} \right] \delta(\tau - \tau_{k}) \tag{1.69}
$$

where $\tau^{*} = -\tau$ to indicate that the adjoint system is solved in reversed time (from end to beginning of interval). The adjoint system requires final and boundary conditions. They are set

$$
\begin{aligned}
\lambda(\tau_{end}) &= 0 \\
\lambda_{\Omega} &= 0
\end{aligned}
\tag{1.70}
$$

The convenience of conditions (1.69-1.70) is seen by substitution in (1.68). Given that the control variables are independent, the condition at the minimum reads

$$
\begin{aligned}
C_{\alpha}^{-1}(\alpha-\alpha_{prior})-\left[\frac{\partial M}{\partial \alpha}\right]^{\mathrm{T}}\lambda=0\\
C_{b}^{-1}(\chi_{\Omega}-\chi_{b})-\left[\frac{\partial M}{\partial \chi_{\Omega}}\right]^{\mathrm{T}}\lambda=0\\
Q^{-1}\varepsilon-\left[\frac{\partial G}{\partial \varepsilon}\right]^{\mathrm{T}}\lambda=0\\
-\lambda(\tau-\Delta\tau)+C_{f}^{-1}(\chi_{\tau-\Delta\tau}-\chi_{prior})=0
\end{aligned}
\qquad (1.71)
$$

where λ is solution of (1.69-1.70). It is critical to notice that this solution depends on χ. Consequently, λ in (1.71) is the solution of (1.69) for $\chi_{\min}$ only. This property implies that the dual system (1.55) and (1.69) still must be solved simultaneously to make use of (1.71)

The way out of this difficult condition which is virtually impossible to overcome in practice with the large multidimensional models in the Geosciences, is to observe that the solution of (1.69) for an arbitrary point χ is directly proportional to the directional gradient of cost function at the same point. This is easily seen from (1.71) but without setting $\delta F=0$. For example

$$
\delta F_{\chi(\tau-\Delta\tau)}=\delta\chi_{\tau-\Delta\tau}^{\mathrm{T}}\frac{\partial F}{\partial \chi_{\tau-\Delta\tau}}=\delta\chi_{\tau-\Delta\tau}^{\mathrm{T}}\left(-\lambda_{\tau-\Delta\tau}+C_{f}^{-1}(\chi_{\tau-\Delta\tau}-\chi_{prior})\right)
$$

when prior is neglected, for simplicity, the resulting relationship reads

$$
\frac{\partial S}{\partial \chi_{\tau-\Delta\tau}}=-\lambda_{\tau-\Delta\tau} \qquad (1.72)
$$

Bennett, A., 1992: Inverse Methods in Physical Oceanography. *Cambridge University Press.* pp 346.

Evensen, G., 2006: Data Assimilation: The Ensemble Kalman Filter. *Springer,* pp

Jazwinski, H., 1970: Stochastic Processes and Filtering Theory. *Mathematics in science and Engineering, Vol. 64. Academic Press.* pp 376.

Kalman, R.E. (1960). A new approach to linear filtering and prediction theory, *Journal of Basic Engineering, Transactions ASME, Series D,* 82, 35-45.

Kalnay, E., 2003: Atmospheric Modeling, Data Assimilation and Predictability. *Cambridge University Press.* pp 341.

Lewis., J., S. Lakshmivarahan and S. Dhall, 2005: Dynamic Data Assimilation: A least squares problem. *Cambridge University Press.* pp 680.

Mosegaard, K., and, A. Tarantola, 2002: Probabilistic Approach to Inverse Problems. *International Handbook of Earthquake & Engineering Seismology (Part A), Academic Press,* 237-265.

O'Neill, A., P.-P.. Mathieu, and, C. Zehner, 2004: Making the most of Earth observation with data assimilation. *ESA Bulletin ,* No. 118, p. 32 – 38.

Rodgers, C. D., 2000: Inverse Methods for Atmospheric Sounding : Theory and Practice. *World Scientific Pub Co Inc* . (*Series on Atmospheric Oceanic and Planetary Physics)*. pp

Tarantola, A., 2005: Inverse Problem Theory and methods for model parameter estimation. *SIAM, Philadelphia,* pp 342.

van Leeuwen, P. J., 2003: A variance-minimizing filter for large-scale applications. *Mon. Wea. Rev.*,131, 2071-2084.