AT721 Section 12:

Introduction to Bayes theorem and General Linear inverse

General references:
Rodgers, C.D., 2000; Inverse Methods for atmospheric sounding, *World Sci*
Bernardo and Smith, 2000; Bayesian Theory, Wiley, 586pp
Other References:
Evans et al., 2002; Submillimeter-wave cloud ice radiometer: Simulations of retreival algorithm performance, *J. Geophys. Res.*, 107, 10.1029/2001JD000709
L'Ecuyer and Stephens, 2002; An Uncertanty model Bayesion Monte carlo retrieval algorithms: application to the TRMM observing system, *Q.J.Roy. Meteorol. Soc.*, 128, 1713-1737.
Strand, 1974; coefficient errors caused by using the wrong covariance matrix in the general linear model, *Annal Statistics*,2, 935-949.

We have seen that the solutions to many inversion problems require compromises between the kind of information wanted and the kind of information that in fact can be derived from any given data set. These compromises are forced upon us when we attempt to introduce information into the inversion process that is not known precisely but rather is known only within some bounds.

12.1 Probability and pdfs

A probability represents a state of knowledge rather than a physical entity. Add discussion on probability

Suppose x is a variable that takes on a continuous set of values over some interval on the real axis, $a \leq x \leq b$. We further assume there exists a piecewise continuous function $p_X(x)$ such that the probability, $P(a \leq x \leq b)$, that X has a value in the interval $a \leq x \leq b$ is given by the area under the curve $p_X(x)$ between $x = a$ and $x = b$

$$P(a \leq x \leq b) = \int_a^b p_X(x)dx \tag{12.1}$$

where $p_X(x)$ is the probability density function (or pdf). The pdf must satisfy

$$p_X(x) \geq 0$$

$$\int p_X(x)dx = 1$$

where the interval is over the entire range of x.

12.1.1 Properties of pdfs

A pdf is a continuous function that may have a complicated shape. Therefore it is helpful to derive a few quantities from the distributions that summarize its major properties. One useful property is some indication of the most likely measurement which is the one of highest probability (Fig. 12.1) occurring at the peak of the distribution , referred to as the maximum likelihood point. If the distribution is skewed, this value, however, may not be a good indication of the most typical value. In these circumstances, the expected or mean value is a better characterization,

$$\langle x\rangle = \int x p_X(x)dx \tag{12.2}$$

where $\langle x\rangle$ is the expected or mean value of x. The nth moment of x is defined thus follows

$$\langle x^n\rangle = \int x^n p_X(x)dx = 1 \tag{12.3}$$

and the variance of x is $[\langle x^2\rangle - \langle x\rangle^2]$ and the standard deviation

$$\sigma_x = [\langle x^2\rangle - \langle x\rangle^2]^{\frac{1}{2}} \tag{12.4}$$

which characterizes the width of the distribution.

Fig. 12.1 The maximum likelihood point of the pdf for data gives the most probable value of the data. In general, this value can differ from the mean datum $\langle x\rangle$.

12.1.1 Joint Probability and correlated data

Experiments usually involve collection of more than one datum. We therefore need to quantify the probability that a set of random numbers take on a given value. For example, for continuous random variables x and y, the joint probability is the probability that both x and y fall in specified ranges and

$$\int\int p_{XY}(x,y)dxdy = 1 \tag{12.5}$$

Furthermore, the marginal pdf for X is obtained by integrating over y

$$p_X(x) = \int p_{XY}(x,y)dy \tag{12.6}$$

In some cases, the data are correlated. High values of one datum, for example, occur either with high or low values with another datum (Fig. 12.2a). The joint distribution takes this correlation into account and given this distribution, correlations can be tested by selecting a function that divides the x, y plane into four quadrants of alternating sign centered on the center of the distribution (Fig. 12.2b). Correlated distributions tend to be concentrated in two opposite quadrants. If $[x - \langle x \rangle][y - \langle y \rangle]$ is the function then the resulting measure of the correlation is the covariance of x and y is defined as

$$cov(x, y) = \int \int [x - \langle x \rangle][y - \langle y \rangle] p_{XY}(x, y) dx dy \tag{12.7}$$

$$= \langle xy \rangle - \langle x \rangle \langle y \rangle$$

which characterizes the basic shape of the joint probability.

Fig. 12.2 (a) The pdf p(x,y) is contoured as a function of x and x. The angle θ is a measure of the correlation and is related to the covariance. (b) The function divides the x-y plane into four quadrants of alternating sign.

The correlation of x and y is defined as

$$cor(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y} \tag{12.8}$$

The correlation $cor(x, y)$ is dimensionless and has the following properties

(i) $cor(x, y) = cor(y, x)$

(ii)$-1 \leq cor(x, y) \leq 1$

(iii) $cor(x, x) = 1, cor(x, -x) = -1$

(iv) $cor(ax + b, cy + d) = cor(x, y)$ if $a, c \neq 0$

For independent x and y, the following hold:

(i) $p_{XY}(x, y) = p_X(x) p_Y(y)$(12.9)

(ii) $\langle xy \rangle = \langle x \rangle \langle y \rangle$

(iii) $\langle (x + y)^2 \rangle - \langle x + y \rangle^2 =$

(iv) $cov(x, y) = 0$ (note the converse of (iv) is not always true).

12.1.1 Functions of random variable

Often problems require the pdf not of x but of some new variable $y = f(y)$, where $f(x)$ is a known function of x. The pdf of y

$$p_y(y) = \int \delta(y - f(x)) p_x(x) dx \tag{12.11}$$

where $\delta(y - f(x))$ is the Dirac delta function.

12.2 The Gaussian Distribution

Although the pdf of many types of variables in many problems may be governed by distributions functions of varying forms, the Gaussian or normal distribution describes the pdfs of a vast number types of problems. In fact, Jaynes points out that if the statistics of a given problem are not known precisely, assumptions for the pdf other than Gaussian introduces spurious information into the problem. We will consider only this form of distribution. Mention also maximum entropy - minimum information of normal dist (Rodgers, chapt 2).

The normalized form of the distribution is

$$p_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[(x - \langle x \rangle)/2\sigma^2] \tag{12.12}$$

where $\langle x \rangle$ and σ are defined above.

12.2.1 Confidence Limits

The probability

$$P(a \le x \le b) = \int_a^b p_X(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b \exp[(x - \langle x \rangle)/2\sigma^2] \tag{12.13}$$

and for the case with $x = 0, a = -\sigma$ and $b = \sigma$, the probability that a parameter falls within the range is 68%. Figure 11.3. presents this probability as a function of parameter range measured in units of σ. For example, the probability that a parameter falls within the range $-2\sigma \le x \le 2\sigma$ is 95.4% and so on.

12.2.2 Joint Probability

The joint distribution of two independent Gaussian variables is, according to 12.x, the product of two Gaussian distributions. When the data are correlated, the distribution has the form

$$p_y(\mathbf{y}) = \frac{1}{2\pi^{\frac{1}{2}} \mid \mathbf{S}_y \mid^{\frac{1}{2}}} \exp[(\mathbf{y} - \langle \mathbf{y} \rangle)^T \mathbf{S}_y^{-1} (\mathbf{y} - \langle \mathbf{y} \rangle)] \tag{12.14}$$

where $\mathbf{S}_y$ is a matrix of correlations between the different data,

$$S_{y,ij} = \int (y_i - \langle \mathbf{y} \rangle)(y_j - \langle \mathbf{y} \rangle) p_Y(y) dy \tag{12.15}$$

When the individual 'measurements' y_i are uncorrelated with respect to measurements y_j then

$$\mathbf{S}_y = \begin{pmatrix} \sigma_1^2 & 0 & \dots \\ 0 & \sigma_2^2 & \dots \\ & \dots & \sigma_n^2 \end{pmatrix}$$

fix up and show examples

12.3 Bayes Theorem

About three centuries ago, people started to give serious thought to the question of how to reason in situations that lack any certainty. Reverand Thomas Bayes is credited with providing an approach that was communicatd after his death in 1761 by Richard Price to the Royal Society in 1763. The technical result at the hear of the essay *An essay towards solving a problem in then doctrine of chances* is what we know know as *'Bayes' theorem'*. It was Laplace in 1812, however, - apparently unaware of Bayes' work - who stated the theorem in its general form that we use today.

The use of Bayes theorem in the context of inverse problems is now widespread. Discussion of this theorem in the context of inverse problems may be facilitated with the the following definitions:

- $p(\mathbf{x})$ as the *prior* pdf of the state $\mathbf{x}$. This means that the quantity $p(\mathbf{x})d\mathbf{x}$ is the probability such that $\mathbf{x}$ lies in the multidimensional volume $(\mathbf{x}, \mathbf{x} + d\mathbf{x})$ expressing quantitatively our knowledge of $\mathbf{x}$ before the measurement is made.
- $p(\mathbf{y})$ as the *prior* pdf of the measurement with a similar meaning. This is the pdf of the measurement before it is made.
- $p(\mathbf{x}, \mathbf{y})$ as the joint prior pdf of $\mathbf{x}$ and $\mathbf{y}$ meaning that $p(\mathbf{x}, \mathbf{y})d\mathbf{x}d\mathbf{y}$ is the probability that $\mathbf{x}$ lies in $(\mathbf{x}, \mathbf{x} + d\mathbf{x})$ and $\mathbf{y}$ lies in $(\mathbf{y}, \mathbf{y} + d\mathbf{y})$
- $p(\mathbf{y} \mid \mathbf{x})$ as the conditional pdf of $\mathbf{y}$ given $\mathbf{x}$ meaning that $p(\mathbf{y} \mid \mathbf{x})d\mathbf{y}$ is the probability that $\mathbf{y}$ lies in $(\mathbf{y}, \mathbf{y} + d\mathbf{y})$ when $\mathbf{x}$ has a given value. This is a function derived by the forward model.
- $p(\mathbf{x} \mid \mathbf{y})$ as the conditional pdf of $\mathbf{x}$ given $\mathbf{y}$ meaning that $p(\mathbf{x} \mid \mathbf{y})d\mathbf{x}$ is the probability that $\mathbf{x}$ lies in $(\mathbf{x}, \mathbf{x} + d\mathbf{x})$ when $\mathbf{y}$ has a given value. This is the quantity of interest for solving the inverse problem.

Fig. 12.3 Fig. 12.3 Illustration of Bayes' theorem for a two-dimensional case.

Figure 12.3 provides a conceptual illustration of the concepts behind these definitions in the context of x and y are scalar. Shown are the contours of $p(x, y$ as well as $p(x)$ and $p(y)$ where

$$p(x) = \int_{-\infty}^{\infty} p(x, y)dy \quad \text{and} \quad p(y) = \int_{-\infty}^{\infty} p(x, y)dx$$

Bayes theorem states that

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \tag{12.16}$$

where the left hand side is the *posterior* pdf of the state when the measurement is given. This represents an updating to our prior knowledge $p(\mathbf{x})$ given the measurement. The most likely value of $\mathbf{x}$ derived from this *posterior* pdf might therefore 'contain' our inverse 'solution'. Our knowledge contained in $p(\mathbf{y} \mid \mathbf{x})$ is explicitly expressed in terms of the forward model and the statistical description of both the error of this model and the error of the measurement. The factor $p(\mathbf{y})$ can practically be ignored as it merely represents a normalizing factor being entirely independent of $\mathbf{x}$.

The general approach to inversion thus follows these basic steps:

- We express our prior knowledge of $\mathbf{x}$ as a pdf
- The measurement process is expressed as a forward model, $f(\mathbf{x})$, which maps the state space $\mathbf{x}$ into a measurement space, via

$$\mathbf{y} = f(\mathbf{x}) + \varepsilon$$

- Bayes' theorem then provides the formalism to invert this mapping and calculate the *posterior* pdf by updating the prior pdf with the measurement pdf
- The most probable solution is then inferred from this posterior pdf

Bayes theorem is general, it is not just a specific inverse method which produces a solution but rather encompasses all inverse methods by providing a way of characterizing all possible solutions and assigning a probability density to each. The forward model is not explicitly inverted and one can obtain different explicit 'answers' to the inverse problem according to how the most 'likely' solution is to be defined. The method provides us with a deeper intuition about how the measurement improves our knowledge of the state.

12.4 An example application

From Frank's work

12.5 Linear inverse problem with Gaussian Statistics

The Bayesian approach provides us with a framework for understanding the inverse problem. Given a measurement together with a description of its error statistics, a forward model describing the relationship between the measurement and the unknown state, and any other a priori information, Bayes relationship allows us to identify the class of possible states consistent with all available information and assign a pdf to them.

One approach to evaluate (12.16) to obtain the *posterior* pdf and thus the most probable solution is to carry out forward model calculations over the entire possible range of $\mathbf{x}$ to establish $p(\mathbf{x} \mid \mathbf{y})$ in the form of a histogram. We have already seen Franks example

Another more explicit way of evaluating of (12.16) follows by invoking the assumption of Gaussian statistics. It follows from (12.14) that

$$-2 \ln p(\mathbf{y} \mid \mathbf{x}) = (\mathbf{y} - f(\mathbf{x}))^T \mathbf{S}_y^{-1} (\mathbf{y} - f(\mathbf{x})) + c_1 \tag{12.17}$$

where c_1 is a constant and $\mathbf{S}_y$ is the measurement error covariance. If we also invoke these statistics to describe our prior knowledge

$$-2 \ln p(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_a)^T \mathbf{S}_a^{-1} (\mathbf{x} - \mathbf{x}_a) \tag{12.18}$$

where $\mathbf{x}_a$ is the a priori value of $\mathbf{x}$ and $\mathbf{S}_a$ is the associated covariance matrix. Combining (12.17) and (12.18) into (12.16) provides the following for the *posterior* pdf:

$$-2 \ln p(\mathbf{x} \mid \mathbf{y}) = (\mathbf{y} - \mathbf{K}\mathbf{x})^T \mathbf{S}_y^{-1} (\mathbf{y} - \mathbf{K}\mathbf{x}) + (\mathbf{x} - \mathbf{x}_a)^T \mathbf{S}_a^{-1} (\mathbf{x} - \mathbf{x}_a) + c_1 \tag{12.19}$$

where we introduce a linear form $f(\mathbf{x}) \approx \mathbf{Kx}$ for the forward model. Since the combination of Gaussian pdfs results in a pdf of the same form, we can write (12.19) as

$$-2 \ln p(\mathbf{x} \mid \mathbf{y}) = (\mathbf{x} - \langle \mathbf{x} \rangle)^T \mathbf{S}^{-1} (\mathbf{x} - \langle \mathbf{x} \rangle) \tag{12.20}$$

where the desired *posterior* pdf is of Gaussian form with the expected value $\langle \mathbf{x} \rangle$ and associated covariance $\mathbf{S}_x$. We can match like terms in (12.20) and (12.19) to arrive at

$$\mathbf{S}_x^{-1} = \mathbf{K}^T \mathbf{S}_y^{-1} \mathbf{K} + \mathbf{S}_a^{-1}$$

and

$$\langle \mathbf{x} \rangle = \mathbf{x}_a + \mathbf{S}_a \mathbf{K}^T (\mathbf{K} \mathbf{S}_a \mathbf{K}^T + \mathbf{S}_y)(\mathbf{y} - \mathbf{Kx}) \tag{12.21}$$

12.6 The maximum a posteriori solution

As noted above, the Bayes relationship allows us to identify the class of possible states consistent with all available information we have to solve our inverse problem. It assigns not a single state but a pdf to represent the class of states. Practical problems require the identification of just one possible state assigning it as a 'solution'. There are a number of possible ways a solution can be constructed from the pdf. Perhaps the most straight forward way of selecting this single state from all states represented by the pdf is to choose either the most likely state, i.e. one for which $p(\mathbf{x} \mid \mathbf{y})$ is a maximum or the expected value solution

$$\langle \mathbf{x} \rangle = \int \mathbf{x} p(\mathbf{x} \mid \mathbf{y}) d\mathbf{x}$$

and for either case, the width of the distribution is a measure of the error. For Gaussian density functions the two solutions are identical and are given by (12.21). For these statistics these solutions are also the same as the minimum variance solutions.

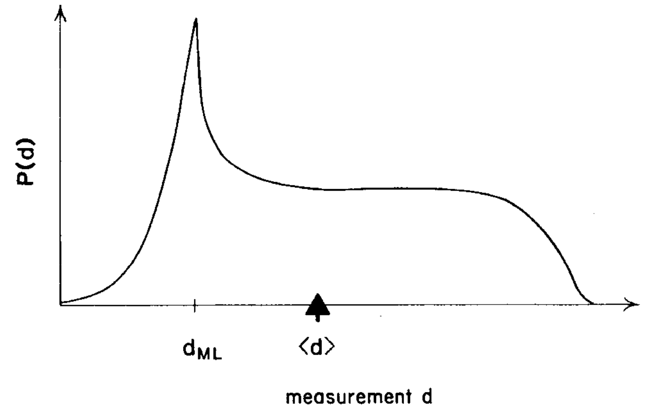


Fig. 12.1 The maximum likelihood point of the pdf for data gives the most probable value of the data. In general, this value can differ from the mean datum <x>.

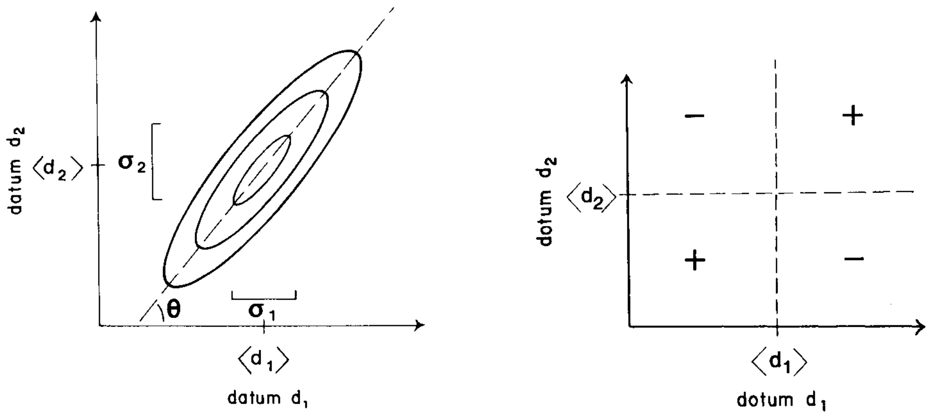
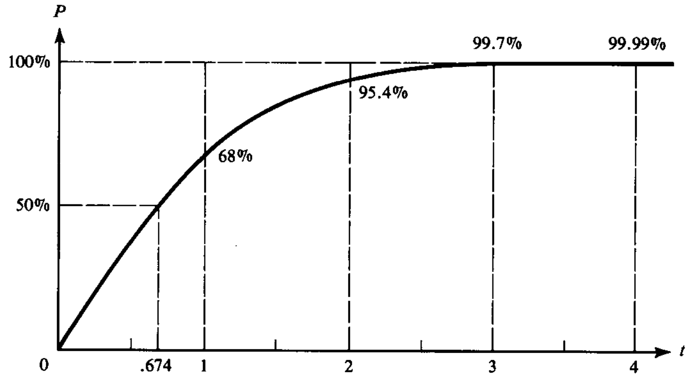


Fig. 12.2 (a) The pdf p(X,Y) is contoured as a function of X and Y. The angle θ *is a measure of the correlation and is related to the covariance. (b) The function* $[\mathrm{x} - \langle \mathrm{X} \rangle][\mathrm{y} - \langle \mathrm{Y} \rangle]$ *divides XY plane into four quadrants of alternating sign.*



t	0	.25	.5	.75	1.0	1.25	1.5	1.75	2.0	2.5	3.0	3.5	4.0
P(%)	0	20	38	55	68	79	87	92	95.4	98.8	99.7	99.95	99.99

Fig. 12.3 The probability that a parameter x falls within t standard deviations of the mean value X.

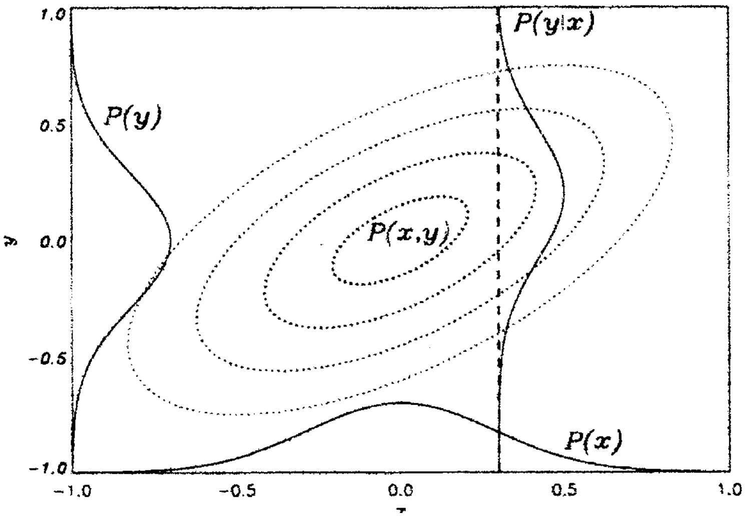


Fig. 12.4 Illustration of Bayes' theorem for a two-dimensional case