

APPLICATIONS OF BAYES THEOREM

Bayes' theorem provides a means for updating previously existing or *a priori* estimates to account for new information or, alternatively, to combine information from different sources. The updated estimates are referred to as *a posteriori* estimates, although in a sequential application of Bayes' theorem to an estimation problem, these *a posteriori* values may in turn play the role of *a priori* estimates for the next iteration of the process.

As presented in an earlier lecture and reviewed here, Bayes' theorem arises directly out of fundamental relationships between conditional and joint probabilities and no one disputes the mathematical correctness of the theorem. However, there is some controversy surrounding application of Bayes' theorem to statistical estimation and decision problems, with statisticians tending to split between "frequentists", who espouse presumably objective approaches to deriving parameter estimates from data, and "Bayesians", who argue that every statistical inference involves the use of fundamentally subjective *a priori* probabilities. Although frequentist approaches are often described as "classical", Bayesian approaches to statistical inference actually predate frequentist approaches. As described by Fienberg (2006), Bayesian techniques have a roughly 250-year history, starting with the Rev. Thomas Bayes' posthumously published paper "An Essay Towards Solving a Problem in the Doctrine of Chances" (1763), if not earlier. Frequentist approaches, on the other hand, really did not exist before the 20th century. Not surprisingly, the adjective "Bayesian" did not exist until frequentists introduced it as a label for *the other guys*.

Bayes' Theorem Using Discrete Probabilities

To review a basic relationship between joint and conditional probabilities, let's use $P(B)$ to represent the probability of occurrence of event B , $P(A, B)$ to represent the joint probability of the simultaneous occurrences of events A and B , and $P(A|B)$ to represent the probability of the occurrence of A given that event B has in fact occurred. The relationship between these probabilities is

$$P(A, B) = P(A|B)P(B).$$

If A and B are independent, then the occurrence of B tells us nothing about the occurrence of A and $P(A|B) = P(A)$, so that the above equation reduces to $P(A, B) = P(A)P(B)$, the joint probability for two independent events. However, if A and B are dependent, then the fact that B has occurred changes our expectations regarding the probability of A , and we have to take this into account in our computation of the joint probability. It is clear that we can also write the joint probability as

$$P(A, B) = P(B|A)P(A).$$

Equating these two expressions for the joint probability yields a relationship between the two conditional probabilities, $P(A|B)$ and $P(B|A)$:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

This equation gives us a means for deriving one form conditional probability from the other. Quite often B represents some underlying *model* or *hypothesis* that is not directly observable and A represents an observable consequence or set of *data* (Sivia, 1996; Aster *et al.*, 2005), so that, ignoring the normalizing factor in the denominator, we can write

$$P(\text{model}|\text{data}) \propto P(\text{data}|\text{model})P(\text{model})$$

where \propto means “is proportional to”. In other words, Bayes’ theorem lets us turn a statement regarding a forward problem (the likelihood of obtaining an observed set of data from a given model) into a statement regarding the corresponding inverse problem (the likelihood that a certain model gave rise to the data we observed), as long as we are willing to make some guesses regarding the probability of occurrence of that model (perhaps among a set of competing models) prior to taking the data into account. This is a valuable relationship because it is usually easier to develop an explicit formulation of the forward problem than of the corresponding inverse problem.

To extend the above relationship, assume that B_i represents one of n possible mutually exclusive events and that the conditional probability for the occurrence of A given that B_i has occurred is $P(A|B_i)$. In this case, the total probability for the occurrence of A is

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

and the conditional probability that event B_i has occurred given that event A has been observed to occur is given by

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^n P(A|B_j)P(B_j)} = \frac{P(A|B_i)P(B_i)}{P(A)}.$$

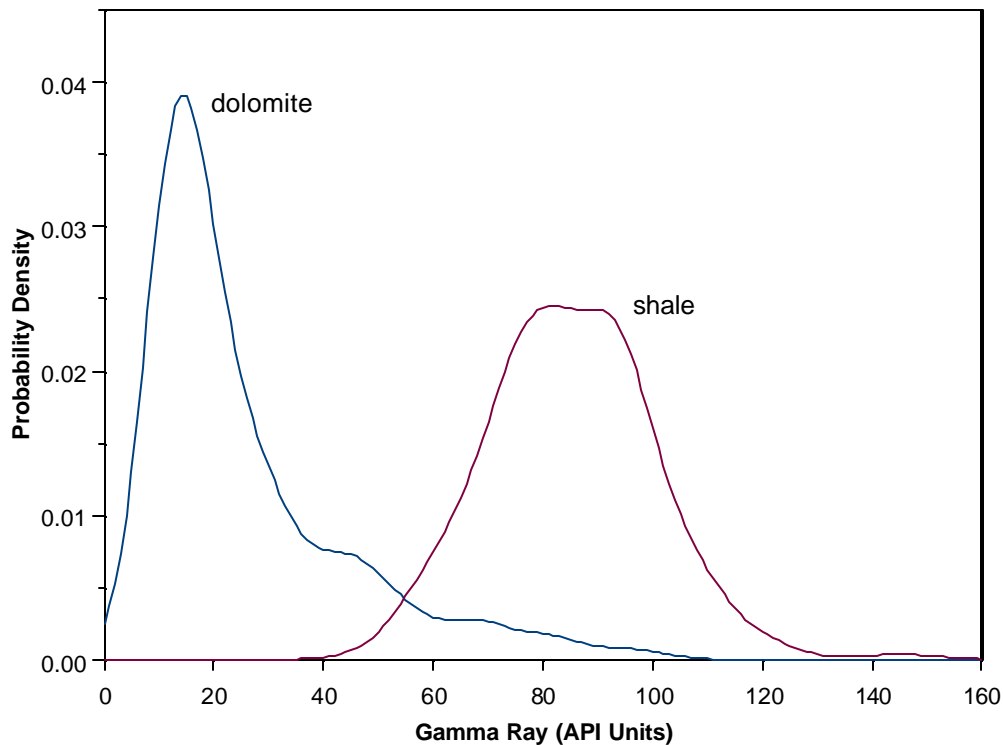
That is, if we assume that event A arises with probability $P(A|B_i)$, from each of the underlying “states” B_i , $i=1, \dots, n$, we can use our observation of the occurrence of A to update our *a priori* assessment of the probability of occurrence of each state, $P(B_i)$, to an improved *a posteriori* estimate, $P(B_i|A)$. If all the conditional probabilities $P(A|B_i)$ are equal, then the occurrence of A gives us no additional information about the underlying states B_i , and the posterior probabilities are equal to the prior probabilities. On the other hand, if the all the prior probabilities are equal (one might say “noninformative”), then

the posterior probabilities are proportional to the conditional probabilities for A . In either case, the denominator serves to scale the results to represent a set of probabilities (summing to unity) for the mutually exclusive events $P(B_i|A)$.

As an example, we could use Bayes' theorem in the problem of discriminating pay from non-pay zones in a reservoir based on measured values of the gamma ray log. The reservoir consists primarily of dolomite intervals constituting the pay zones and shale non-pay intervals. Gamma ray values, measured in API units, tend to be significantly higher in the shales due to the presence of natural radioactive isotopes in the clay minerals. Typical gamma ray values for a mid-continent shale would be around 110 API units, compared to 10 to 15 for typical dolomites. However, the shale in this reservoir shows a significantly lower range of gamma ray values, probably due to a high silt content, and some of the dolomites show high gamma ray values, probably due to the presence of uranium. This leads to fairly noticeable overlap between the gamma ray distributions for dolomites and shales.

Core samples from across the field allow us to tie logged gamma ray values to known lithologies for certain zones, allowing us to estimate the distribution of gamma ray values for each lithology:

Gamma ray distributions for dolomite and shale



The distributions shown above were estimated from the gamma ray values for 476 dolomite intervals and 295 shale intervals. The curves shown are probability density estimates derived from the data using a smoothing kernel; you can think of them as smoothed histograms. We will use these distributions, developed from intervals with known lithology, to make lithology assignments based on gamma ray values in uncored wells.

Clearly, the two gamma ray distributions show some overlap, but we could do a reasonable job of discriminating shale from dolomite by simply assigning intervals with a gamma ray value greater than some threshold value, say 60 API units, to the category “shale” and the rest to the category “dolomite”. Of the 476 known dolomite intervals, 34, or about 7%, show gamma ray values higher than 60. Of the 295 shale samples, 280, or about 95%, are associated with gamma ray values greater than 60. These values give us estimates of $P(A|B_i)$, that is the probability of $\text{GammaRay} > 60$ in dolomite and shale intervals, respectively. Furthermore, if we assume that the proportion of dolomite and shale intervals in our 771 core samples reflects the overall prevalence of the two lithologies in the reservoir, then we could estimate prior probabilities of roughly 60% (476 of 771) for the occurrence of dolomite and 40% (295 of 771) for the occurrence of shale. Thus, our events and probabilities would be defined as:

A :	$\text{GammaRay} > 60$
B_1 :	occurrence of dolomite
B_2 :	occurrence of shale
$P(B_1)$:	prior probability for dolomite based on overall prevalence = 60%
$P(B_2)$:	prior probability for shale based on overall prevalence = 40%
$P(A B_1)$:	probability of $\text{GammaRay} > 60$ in a dolomite = 7%
$P(A B_2)$:	probability of $\text{GammaRay} > 60$ in a shale = 95%

It is important to note that the conditional probabilities $P(A|B_1)$ and $P(A|B_2)$ do not represent a set of mutually exclusive events, but instead are characteristics of the distributions of gamma ray values for dolomites and shales, respectively.

In this case, the total probability for the occurrence of a gamma ray value greater than 60 is

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) = 0.07 * 0.60 + 0.95 * 0.40 = 0.422.$$

If we measure a gamma ray value greater than 60 at a certain depth in a well, then the probability that we are logging a dolomite interval is

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A)} = \frac{0.07 * 0.60}{0.422} = 0.10$$

and the probability that we are logging a shale interval is

$$P(B_2|A) = \frac{P(A|B_2)P(B_2)}{P(A)} = \frac{0.95 * 0.40}{0.422} = 0.90.$$

Thus, our observation of a high gamma ray value has vastly altered our assessment of the probabilities of occurrence of dolomite and shale from 60% and 40%, based on our prior estimates of overall prevalence, to 10% and 90%.

Of course, our prior estimates involve a fair amount of subjectivity and our posterior estimates would change somewhat if we used different priors. However, given the strong contrast in the probabilities of measuring a high gamma ray value in shales versus dolomites, an observed gamma ray value greater than 60 would still lead us to decide that the measured interval is most likely a shale, even under a fairly wide variation in prior estimates. For example, if we used prior estimates of 80% and 20% for the occurrence of dolomite and shale, respectively, making it more likely that we will decide in favor of dolomite, the high gamma ray value would still lead us to a posterior probability of 77% for the occurrence of shale in the logged interval.

Bayes' Theorem Using Probability Densities

It is also possible to formulate Bayes' theorem using probability density functions in place of the discrete probabilities $P(A|B_i)$. That is, we may be measuring a continuous variable, X , that follows a different probability density function for each underlying category, B_i . In our example above, the continuous variable would be gamma ray and the two categories would be dolomite and shale. We could represent the probability density function that X follows in each case as $f(x|B_i)$ or, more compactly, $f_i(x)$. The continuous-variable rendition of Bayes' theorem is then written as

$$P(B_i|x) = \frac{f_i(x)P(B_i)}{\sum_{j=1}^n f_j(x)P(B_j)}$$

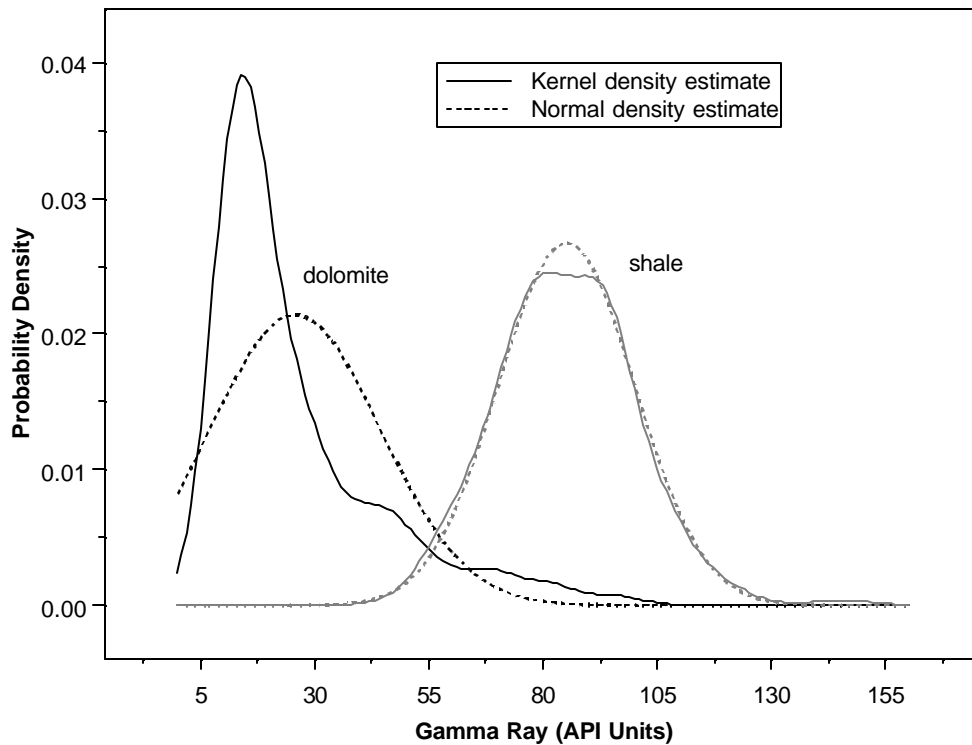
That is, if we can characterize the distribution of X for each category, B_i , we can use the above equation to compute the probability that event B_i has occurred given that the observed value of X is x . For example, based on the observed distribution of gamma ray values for dolomites and shales, a gamma ray measurement of 110 API units almost certainly arises from a shale interval.

We can use this version of Bayes' theorem to develop a continuous mapping from gamma ray value to posterior probability, eliminating the need to choose a threshold gamma ray value. To do so, we must find some means for estimating the probability density function for each category. We could employ a normal density function, using the sample mean and standard deviation of the gamma ray values for the intervals of known lithology to estimate these parameters:

	Dolomite	Shale
Mean	25.8	85.2
Std. Dev.	18.6	14.9
Count	476	295

The resulting normal distribution for shale represents the data distribution quite well, except for the flattened peak, but the normal estimate for the dolomite distribution is fairly poor, due primarily to the skewness of the observed distribution:

Normal Approximations for Gamma Ray Distributions



We can plug the estimated gamma ray means, \bar{x}_1 and \bar{x}_2 , and standard deviations, s_1 and s_2 , into the normal density function to obtain

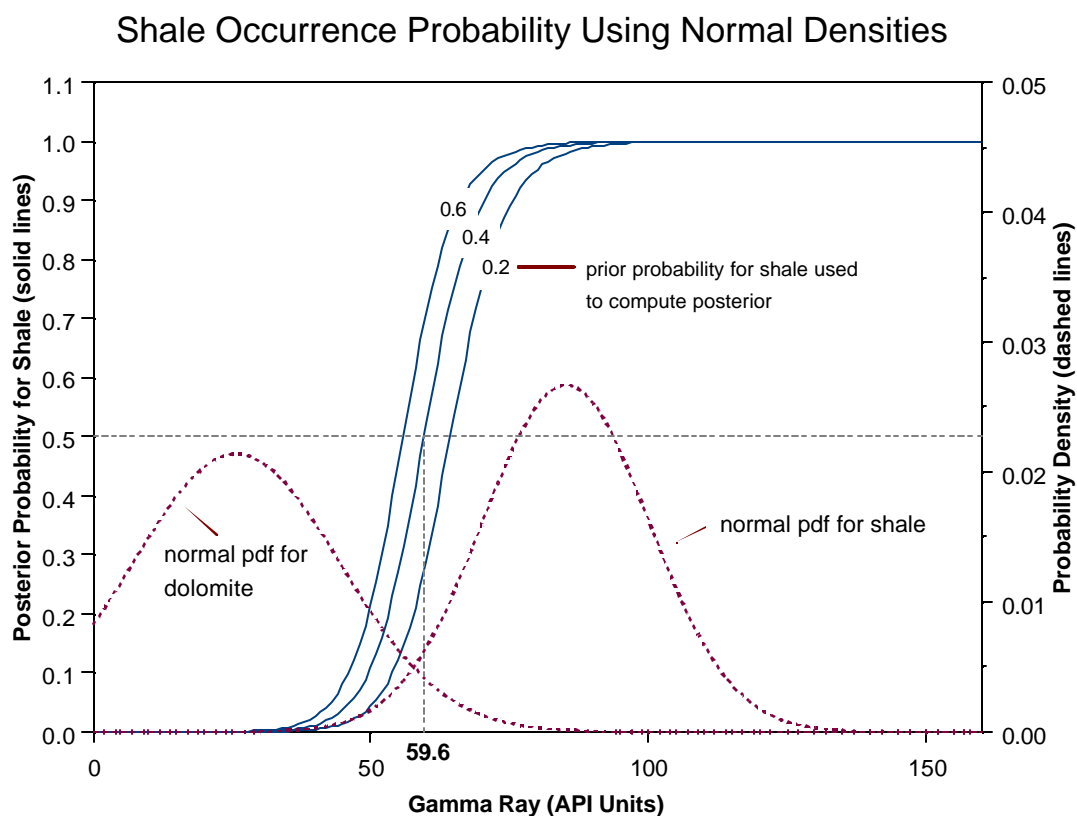
$$f_1(x) = \frac{1}{s_1 \sqrt{2\pi}} \exp\left[-(x - \bar{x}_1)^2 / 2s_1^2\right]$$

and

$$f_2(x) = \frac{1}{s_2 \sqrt{2\pi}} \exp\left[-(x - \bar{x}_2)^2 / 2s_2^2\right]$$

for dolomite (1) and shale (2), respectively.

Substituting these expressions, along with estimated values for the prior probabilities, yields a formula for $P(B_i|x)$ as a continuous function of x . In this example we could compute the posterior probability for the occurrence of shale and plot this versus observed gamma ray value, perhaps plotting curves for different prior probabilities of shale to judge the sensitivity to this parameter:

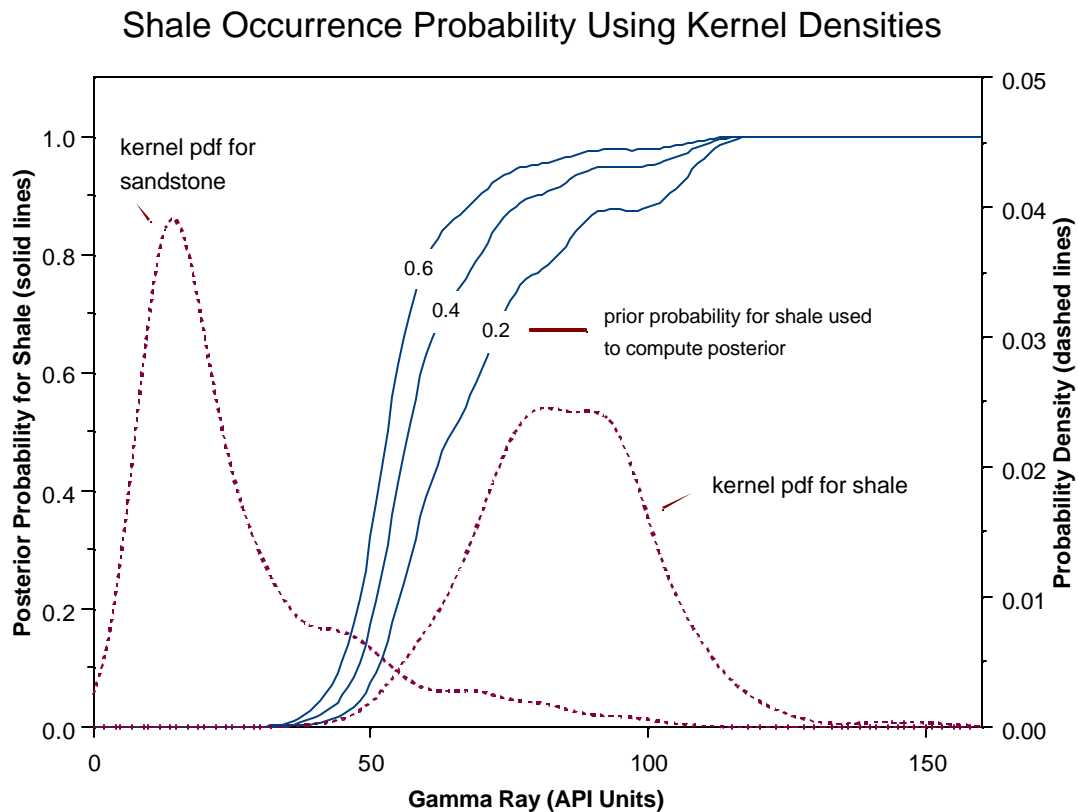


In this two-category example, the posterior probability for the occurrence of dolomite is the complement of (that is, 1 minus) the posterior probability for shale, so we needn't plot both curves, and the prior probabilities are also complementary, so we only need to look at the sensitivity with respect to one of the priors. In this case, Bayes' theorem combines the two density functions, weighted according to the prior probabilities, to compute an S-shaped curve representing the probability for the occurrence of shale versus measured gamma ray value – essentially, a “soft” threshold function. Typically, one would assign an interval to the lithology with the highest posterior probability for the measured gamma ray value, essentially making a hard threshold at the gamma ray value where the posterior probabilities both equal, at 50%. Assigning an observation to the

class with the highest posterior probability is sometimes referred to as *Bayes' rule allocation*.

A higher prior for shale increases the posterior probability for shale at any given gamma ray value, but clearly we would make the same assignment for most gamma ray values over a wide range of priors. For the base case of a 40% prior probability for shale, the 50% posterior probability value occurs at a gamma ray value of about 59.6, so using this point in the curve would lead to essentially the same assignments as the hard threshold at 60 API units that we used in the discrete probability example. What we have gained is the ability to get estimates of the posterior probabilities for any gamma ray value. We could use this, for example, to convert a gamma ray log for a well into a continuous-valued log of shale occurrence probability versus depth.

Just to emphasize the generality of this form of Bayes' theorem, we could carry out the same exercise using the kernel density estimates, rather than the normal density estimates, for the same sequence of gamma ray values, resulting in more interesting but still generally S-shaped curves for the posterior probability of shale occurrence:



Note that the transition range for these posterior probability curves is wider than those based on the normal density functions, since we are now better representing the overlap between the gamma ray distributions for the two lithologies.

We could extend this analysis to multivariate problems simply by using multivariate density functions in Bayes' theorem. That is, \mathbf{x} could represent a vector of measurements on a set of logs, rather than a single log. The application of Bayes' theorem is exactly the same in this case, but the development of the density functions will be more involved.

Applying Bayes' theorem for discrimination using normal density functions leads directly to classical *discriminant analysis* (McLachlan, 1992). In the multivariate case, the distribution of \mathbf{X} for each category would be characterized by the vector mean and covariance matrix computed from observations known to arise from that category. If the covariance matrices for the different classes are all assumed to be equal (estimated by the "pooled" covariance matrix for the groups), then plugging the resulting density functions into Bayes' theorem leads to *linear discriminant analysis*: assigning a given \mathbf{x} value to the category with the highest posterior probability amounts to drawing linear boundaries between categories in variable space. If a different covariance matrix is used for each group, then the boundaries drawn by this Bayes' rule allocation are quadratic and the approach is termed *quadratic discriminant analysis*.

As discussed in McLachlan (1992), Bayes' theorem can be written to take into account *misallocation costs*, c_{ij} , the investigator's assessment of the cost of assigning an observation that has actually arisen from class i to class j . These misallocation costs multiply the prior probabilities in Bayes' theorem, essentially inflating or deflating the priors proportionally and thus expanding or contracting the regions of variable space assigned to each class accordingly. Applying Bayes' rule allocation in this case minimizes the overall misallocation cost.

For example, we may be trying to distinguish between several different facies, some "pay" and some "non-pay", on the basis of a set of well logs. In this case, the cost of misallocating one non-pay facies as another non-pay facies or one pay facies as another pay facies would be relatively low, whereas the costs of calling a pay facies a non-pay facies, or vice-versa, would be fairly high. It is quite possible that there would be some asymmetry in our assessment of misallocation costs. If we were particularly averse to missing pay zones, we would probably assign a higher cost to the classification of a pay facies as a non-pay facies than we would to the opposite misclassification. This would expand the region of variable (log) space assigned to the pay facies, relative to the equal-cost scenario. This would reduce the chances of missing pay zones but increase the number of *false positives* – non-pay zones predicted as pay zones. Thus our assigned misallocation costs would relate directly to our estimates of the actual dollar cost of missing a pay zone versus the expense of more detailed investigation of what turn out to be non-pay zones.

Evaluation of the costs of different decisions under different possible scenarios (that is, competing models with varying prior probabilities) is the topic of *statistical decision analysis*, which is covered in some detail for petroleum engineering applications by Harbaugh *et al.* (1995) and for civil engineering by Benjamin and Cornell (1970).

References

- R.C. Aster, B. Borchers, and C.H. Thurber, 2005, *Parameter Estimation and Inverse Problems*, Elsevier Academic Press, 301 pp.
- J.R. Benjamin and C.A. Cornell, 1970, *Probability, Statistics, and Decision for Civil Engineers*, McGraw-Hill Book Company, 684 pp.
- S.E. Fienberg, 2006, When Did Bayesian Inference Become “Bayesian”?, *Bayesian Analysis* 1(1), pp. 1-40, <http://ba.stat.cmu.edu/journal/2006/vol01/issue01/fienberg.pdf>.
- J.W. Harbaugh, J.C. Davis, and J. Wendebourg, 1995, *Computing Risk for Oil Prospects: Principles and Programs*, Pergamon, 464 pp.
- G.J. McLachlan, 1992, *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, Inc., 526 pp.
- D.S. Sivia, 1996, *Data Analysis: A Bayesian Tutorial*, Oxford University Press, 240 pp.