Chapter 4 Bayes' theorem and linear discriminant analysis applied to seismic early warning

The development of Virtual Seismologist (VS) method for seismic early warning is the topic of this chapter. Each of Chapters 5 through 8 is a case study of the VS method applied to the data set of a real Southern California earthquake.

The Virtual Seismologist method is a Bayesian approach to seismic early warning inspired by the relative speed and accuracy with which expert seismologists can estimate magnitude and distance by combining their "background" knowledge of earthquake activity with quick visual analysis of available waveform data. From the shape and relative frequency content of incoming data, expert seismologists are typically able to estimate quickly the magnitude and location of earthquakes. Part of this can be attributed to the experience that (presumably) an expert seismologist would have in analyzing waveform data. However, experience in analyzing waveform data is not the only factor that contributes to the speed and relative accuracy of human seismologists in estimating earthquake source parameters. That is, estimating earthquake source parameters from limited data is not merely a pattern recognition problem (though it may largely be one). A "background" knowledge regarding relative earthquake probabilities is an essential component of how human seismologists process information. Knowledge of fault locations, relative frequency of earthquake magnitudes (Gutenberg-Richter law), spatial and temporal changes in observed seismicity, on-going earthquake sequences, and the state of health of the seismic network monitoring the earthquake activity are some of the disparate types of information relevant to the source estimation problem. The Virtual Seismologist method for seismic early warning uses Bayes' theorem to combine pattern recognition-type approaches on the incoming waveform data with relevant“background” knowledge to estimate quickly the magnitude and location of an on-going earthquake from whatever data are available. (Within the context of this thesis, the term “source estimation problem” means the problem of estimating magnitude and distance or location given the available waveform data, which may be sparse.)

Standard location methods

It is necessary to locate an earthquake before its magnitude can be determined. Locating an earthquake requires determining its hypocentral coordinates and its origin time. Standard location methods require arrivals from at least 3 stations; additional arrivals allow error estimates to be calculated. Given a set of observed arrival times, a location is obtained by minimizing the residual between observed and predicted arrivals. This is a nonlinear problem that is usually solved iteratively (i.e., assume a location and origin time, predict the arrivals at stations with observed arrivals, calculate the residual between observed and predicted arrivals, repeat the process until a minimum in the residuals is found).

Once a location estimate is available, the earthquake magnitude can be determined. Magnitude is usually defined based on peak amplitude of a particular phase (P-wave for body wave magnitude, m_b, Rayleigh wave for surface wave magnitude, M_s, and S-wave for local magnitude, M_L) of ground motion; the exception is the moment magnitude M_w, which is obtained by fitting long period waveforms. In general, the standard approach for locating an earthquake and determining its magnitude is relatively slow for seismic early warning, because it requires arrivals at multiple stations and peak amplitudes, which are typically associated with slower traveling phases. A different approach is necessary for seismic early warning.

When time *is* of the essence: seismic early warning

In seismic early warning, time *is* of the essence. The goal in seismic early warning is to provide short-term warning (on the order of seconds to tens of seconds) of imminent strong ground motion from large earthquakes (Heaton, 1985). Once the earthquake rupture triggers stations within the epicentral region, the idea is to quickly (and reliably) estimate magnitude and location, and use these estimates to predict the expected levels of shaking in areas further from the epicentral region. One possible measure of an early warning system's success would be if the warning information reaches the subscribers to the system before the arrival of the damaging seismic energy (typically the S-wave) at their sites, that is, if the available warning time, T_{warn}, is greater than zero.

$$T_{warn} = T_S - T_{est} - T_{transit} \tag{4.1}$$

where T_{warn} is warning time in seconds, T_S is the S-wave arrival time at the target warning area, T_{est} is the time necessary before estimates of magnitude and location or distance can be obtained, and $T_{transit}$ is the time required for information to travel from the station to the central processing facility, and from the central processing facility to users receiving the warnings. Since information travels much faster than seismic waves through the earth, it is assumed that $T_{transit}$ is small relative to the other quantities in Eqn. 4.1. ($T_{transit}$ might be significant in comparing different methods for seismic early warning, depending on which data stream output from the seismic stations is utilized in the estimation process.) Therefore, to maximize T_{warn}, it is necessary minimize to T_{est}. T_{est} is a function not only of how long the calculations take, but more importantly, how much information from the earthquake rupture is necessary before magnitude and location can be estimated. Standard location methods have relatively large T_{est}, since they require arrivals at multiple stations and peak amplitudes, which are typically measured from slower-traveling phases.

P-wave as information carrier

Estimating magnitude from P-waves provides gains in the available warning time. The first use of P-wave waveforms to estimate magnitude and location is typically credited to Nakamura (1988), who developed UrEDAS (Urgent Earthquake Detection and Alarm System), the early warning system deployed to mitigate earthquake-related damage to the Japanese railway system. In Southern California, Allen and Kanamori (2003) developed the ElarmS (Earthquake Alarm System) for proof of concept on the Southern California Seismic Network (SCSN). The scientific basis for estimating earthquake size from P-wave amplitudes is described by in a paper by Kanamori (2004). The underlying concept (in Kanamori's terminology) is that P-waves are carriers of information, and S-waves carriers of energy. Information regarding the size of the earthquake can be obtained from P-wave frequency content; small earthquakes involve small patches of slip and radiate relatively high-frequency energy compared to larger magnitude events, whose finite rupture dimensions contribute to large energy in the lower frequencies (Allen and Kanamori, 2003). Allen and Kanamori (2003) use the predominant period of the P-wave to estimate magnitude. A new method for magnitude estimation method using ratios of ground motion will be presented in this Chapter; this ratio-based method shares with Allen and Kanamori (2003) and Nakamura (1988) the concept of using relative frequencies of the P-wave to determine magnitude.

4.1 Introduction to Bayes' theorem

The Virtual Seismologist (VS) method is a Bayesian approach to seismic early warning. Prior to a discussion of the advantages of being Bayesian, Bayes' theorem will be reviewed in the context of seismic early warning. The following Section parallels the presentation of introductory Bayesian concepts in "Data Analysis: A Bayesian Tutorial", by Sivia (1996).

From Sivia, the usual rules of probability theory are

$$prob(X|I) + prob(\bar{X}|I) = 1 \tag{4.2}$$

$$prob(X, Y|I) = prob(X|Y, I) \times prob(Y|I) \tag{4.3}$$

where X and Y are two propositions, $\bar{X}$ denotes "not X", the vertical bar "|" denotes "given" or "conditioned upon", and a comma is read as the conjunction "and". Eqn. 4.2 is known as the *sum rule.* The sum rule states that if we specify our degree of belief in the truth of a proposition X, $prob(X|I)$, we are implicitly specifying how much we belief it is false, $prob(\bar{X}|I)$. Eqn. 4.3 is known as the *product rule.* The product rule states that if we specify our belief in the truth of Y, $prob(Y|I)$, and specify our belief in the truth of X given Y, $prob(X|Y,I)$, then we are in fact specifying how much we believe both X and Y are true, $prob(X,Y|I)$. There are no absolute probabilities; everything is conditioned on "I", which can be thought of as "background" information.

Bayes' theorem follows from the product rule. Eqn. 4.3 gives the probability that both X and Y are true. Eqn. 4.4 is equivalent to Eqn. 4.3. The order of X and Y in the left hand side term is switched and the product rule applied.

$$prob(Y,X|I) = prob(Y|X,I) \times prob(X|I) \tag{4.4}$$

Setting the right-hand sides of Eqn. 4.3 and 4.4 equal gives

$$prob(X|Y,I) \times prob(Y|I) = prob(Y|X,I) \times prob(X|I) \tag{4.5}$$

or, after rearranging,

$$prob(X|Y,I) = \frac{prob(Y|X,I) \times prob(X|I)}{prob(Y|I)} \tag{4.6}$$

Eqn. 4.6 is known as *Bayes' theorem.* Its usefulness becomes evident when X is replaced with "hypothesis" and Y with "observed data". In seismic early warning, the primary question of interest is: what magnitude and location (or epicentral distance) estimates are most probable given the limited available observations from the on-going earthquake? (For now, the discussion will be phrased in terms of epicentral distance, R. Similar arguments holds for epicentral location (latitude, longitude) instead of epicentral distance.) In terms of the quantities in Bayes' theorem, the hypothesis is "M is the magnitude of the earthquake, and R is the epicentral distance". The observed data are the (initially limited set of) available observations. The most probable estimates of magnitude and location (or distance) are the M, R for which the probability density function *prob('M is the magnitude of the earthquake', 'R is the epicentral distance (or location)'| 'limited available observations')* (or *prob(hypothesis|data)*) is a maximum. It is not clear how to assign such probabilities directly. Conveniently, Bayes' theorem states that $prob(hypothesis|data)$ is a function of $prob(data|hypothesis)$. The probability density function $prob(data|hypothesis)$ is easier to define. In fact, in seismic early warning, all that is required to define $prob(data|hypothesis)$ is a ground motion model that relates the observed quantities (the available ground motion data) to the parameters of interest (magnitude and epicentral distance). If the observations being considered are envelopes of ground motion, then the envelope attenuation relationships discussed in the previous Chapters are precisely what are need to define the probability density function $prob(data|hypothesis)$. This is the power of Bayes' theorem. It allows us to define the quantity of interest, $prob(hypothesis|data)$, which is hard to define, as a function a quantities which can be defined, $prob(data|hypothesis)$. All that is necessary is a model relating the observations to their causative parameters.

Rewriting Bayes' theorem in terms of “data” and “hypothesis” gives

$$prob(hypothesis|data) = \frac{prob(data|hypothesis) \times prob(hypothesis)}{prob(data)} \tag{4.7}$$

Following Sivia (1996), the conditioning on background information “I” is dropped in notation, but not in concept. Sivia cautions against forgetting these initial assumptions on background information, citing such lapses as the likely cause of debates regarding interpretation of end results of data analysis.

Each of the terms in Eqn. 4.7 is strictly a probability density function. Recall that, given a probability density function, say $prob(X)$, the probability of X taking on a value between x_1 and x_2 is given by

$$Pr(x_1 \leq X < x_2) = \int_{x_1}^{x_2} prob(X)\, dX \tag{4.8}$$

From here on, “Pr” is used to denote actual probabilities, and “prob” to denote probability densities. Each of the probability densities in Eqn 4.7 has a formal name. $prob(hypothesis)$ is called the Bayesian *prior.* It represents the state of knowledge or beliefs regarding the phenomena being studied *before* considering the current data. The data modify these prior beliefs by means of the *likelihood function*, $prob(data|hypothesis)$. Defining the likelihood function, requires a model (or models) relating the hypothesis to the observations. The likelihood function quantifies how well the observed data fit predictions of the ground motion model given certain hypotheses. Finally, the quantity of primary interest, $prob(hypothesis|data)$, is called the Bayesian *posterior.* The posterior represents the state of knowledge regarding the phenomena being studied accounting for both prior beliefs *and* the observations. If the observations are consistent with the prior beliefs, then the posterior should reflect a stronger degree of belief in the prior hypothesis. On the other hand, if the observations are inconsistent with the prior, then the posterior will reflect a lesser degree of belief in the prior hypothesis. The influence of the prior decreases with increasing number of observations. Maximizing the posterior yields the hypothesis that is most consistent with the data. Finally, the denominator in Eqn. 4.7, $prob(data)$, is known as the *evidence.* It plays an important role in model class selection (Beck and Yuen, 2004). However, in this thesis, Bayes' theorem is used primarily in parameter estimation, or finding the 'best' hypothesis given the observations. In parameter estimation applications, the evidence is treated as a normalizing constant, since it is not an explicit function of the hypothesis.

4.2 Bayes' theorem in seismic early warning

Thus far, the discussions regarding "hypotheses" and "observations" have been somewhat vague. In seismic early warning, the question of interest is: given the available observations from an on-going earthquake rupture, what are the most likely or most probable estimates of magnitude and distance (or location) of the earthquake, and how do these estimates evolve as more data become available? In seismic early warning, the "observations" are ground motion amplitudes recorded from the seismic network and the "hypothesis" is "M and R (or latitude, longitude) are the magnitude and location of the earthquake causing the observed ground motion amplitudes". In terms of the quantities involved in seismic early warning, Bayes' theorem is

$$\begin{aligned} prob(M,R|A) &= \frac{prob(A|M,R) \times prob(M,R)}{prob(A)} & (4.9) \\ &\propto prob(A|M,R) \times prob(M,R) & (4.10) \end{aligned}$$

where A is a vector of the available observed ground motion amplitudes. The likelihood, $prob(A|M,R)$, is defined in terms of ground motion models relating magnitude and distance to observed ground motion amplitudes, i.e.,the envelope attenuation relationships developed and discussed in the Chapters 2 and 3. Eqn. 4.9 and 4.10 say that the most likely magnitude and location estimates consistent with the available observed ground motion amplitudes are a function of the expected ground motions given by the ground motion models and the prior beliefs. Bayes' theorem in the form of Eqn. 4.10 can be used since this is a parameter estimation problem, and the evidence, $prob(A)$ is not a function of M and R. In the statement of the seismic early warning problem, the prior, $prob(M,R)$ represents beliefs regarding relative earthquake probabilities in terms of size and location. The degree of complexity that can be incorporated into the prior is very flexible. Little knowledge (via a uniform prior) can be assumed for simplicity. More complex models describing earthquake occurrence can also be included. For example, in seismic early warning, the prior could be uniform over magnitude and distance. While this simplifies calculations, it is an overly conservative representation of the state of knowledge regarding earthquake occurrence. A uniform prior in M and R implies that earthquakes of all magnitudes and at all distance (or locations) are equally likely. This is certainly not the case. It has long been accepted that magnitude-frequency relationships of earthquakes follow the Gutenberg-Richter law. Earthquakes occur on faults, and they often cluster in time and space. Thus, knowledge regarding previously observed seismicity can be extremely relevant prior information. The potential use of prior information afforded by the Bayesian framework is perhaps the single most important distinction between the Virtual Seismologist method and other proposed paradigms for seismic early warning. The types of information that can be useful priors will be discussed later in this chapter.

The following Sections discuss how to define the various terms in the Bayesian statement of the seismic early warning problem (Eqn. 4.10).

4.3 Defining the likelihood, $prob(A|M, R)$

The goal in seismic early warning is to find estimates of magnitude and location M, R that are consistent with the observed ground motion amplitudes A. Consider the situation where only P-wave data from a single station is available. Using the standard earthquake location methods discussed previously, this problem is underdetermined.

To focus attention on the likelihood function $prob(A|M, R)$, assume for now a uniform prior, or $prob(M, R) = constant$. Eqn. 4.10 then becomes

$$prob(M, R|A) \propto prob(A|M, R) \tag{4.11}$$

(Note: Since the attenuation relationships are defined in terms of $Y_A = \log_{10} A$, probability density functions involving the observations A should be rewritten in terms of the log of the observations Y_A. A and Y_A should be interchangeably be understood as the available observed data.) Eqn. 4.11 states that the posterior, $prob(M, R|A)$, which contains the information about the source estimates of M, R, is directly proportional to the likelihood, $prob(A|M, R)$. The likelihood expresses the probability of observing the data, A, given that M is indeed the "true" magnitude and that R is indeed the source-to-station distance. How to formulate the likelihood, which describes how plausible the various ground motion amplitudes are, given the earthquake magnitude and epicentral distance to the station, makes use of the envelope attenuation relationships developed in Chapters 2 and 3.

Given the envelope amplitude attenuation relationships (Eqn. 2.3), which express ground motion amplitudes A as a function of magnitude$M$, distance R, and site, and letting $Y = \log_{10}(A)$, the Bayes likelihood function can be expressed as

$$\begin{aligned} prob(Y|M,R) &= \frac{1}{\hat{\sigma}\sqrt{2\pi}} \exp\left(-\frac{(Y - \bar{Y}(M,R))^2}{2\hat{\sigma}^2}\right) & (4.12) \\ \bar{Y} &= aM - b(R + C(M)) - d\log_{10}(R + C(M)) + e & (4.13) \end{aligned}$$

where (a, b, c_1, c_2, d, e) are the regression coefficients appropriate for the body wave phase (P- or S-wave), direction (vertical or horizontal), component (acceleration, velocity, or displacement) and site (rock or soil), and $\hat{\sigma}$ is the appropriate standard error of regression ($\hat{\sigma}^2$ is the best estimate of the variance σ^2 of the errors). In parameter estimation for seismic early warning, the available observations $Y = \log_{10}(A)$ are given, regression coefficients (a, b, c_1, c_2, d, e) and $\hat{\sigma}$ are known, and the source estimates M, R are unknowns. Eqn. 4.12 states that the observed data $Y = \log_{10}(A)$ are being modeled as a normal random variables with μ given by the envelope attenuation relationship Eqn 2.3 and variance $\sigma^2 = \hat{\sigma}^2$. This holds because the errors ϵ in Eqn. 2.3 are additive and normally distributed, and linear functions of normal variables are themselves normal.

Best estimates $\hat{M}$, $\hat{R}$ given a single observed amplitude

Given the observed ground motion amplitude $Y = \log_{10}(A)$, the best estimates for the magnitude of the causative earthquake and the source-to-station distance, M, R, are given by the $\hat{M}$ and $\hat{R}$ (with $\hat{}$ denoting "estimate") that maximize the posterior density function. From 4.10 (assuming a uniform prior $prob(M,R) = constant$)

$$\begin{aligned} prob(M,R|Y) &\propto prob(Y|M,R) & (4.14) \\ &= \frac{1}{\hat{\sigma}\sqrt{2\pi}} \exp\left(-\frac{(Y - \bar{Y}(M,R))^2}{2\hat{\sigma}^2}\right) & (4.15) \end{aligned}$$

To determine the $\hat{M}, \hat{R}$ that maximize the posterior, it is simpler to work directly with the log-likelihood, $L = \log_e prob(M,R|Y)$

$$
\begin{aligned}
L &= \log_e prob(M,R|Y) && (4.16)\\
&= \log_e\left(\frac{1}{\hat{\sigma}\sqrt{2\pi}}\right) - \frac{(Y - Y(\bar{M},R))^2}{2\hat{\sigma}^2} && (4.17)\\
&= \text{constant} - \frac{(Y - \bar{Y}(M,R))^2}{2\hat{\sigma}^2} && (4.18)
\end{aligned}
$$

Maximizing the posterior $prob(M,R|Y)$ therefore involves minimizing $(Y - \bar{Y}(M,R))^2$. Clearly, the maximum or minimum of L is identified by $(Y - \bar{Y}(M,R)) = 0$, which describes a curve in (M,R) space. This gives one equation in two unknowns, and the source estimation problem is under-determined. There are infinite pairs of (M,R) coordinates that satisfy $(Y - \bar{Y}(M,R)) = 0$; trade-offs between (M,R) *cannot* be resolved by a single observed P-wave amplitude.

Fortunately, we are never in the situation just described, with only a single amplitude observation on a single channel available. Digital seismic stations typically have the capability to perform real-time recursive filtering on-site; if the instrument output is proportional to either acceleration or velocity (as is common for strong motion and broadband sensors, respectively) then recursive filters can be designed to integrate and/or differentiate the output data stream such that ground motion acceleration, velocity, and displacement are available.

Best estimates of $\hat{M}, \hat{R}$ given P-wave acceleration, velocity, and displacement amplitudes

The problem with using attenuation relationships to define the likelihood function is that trade-offs between magnitude and distance cannot be resolved when data is only available from a single station. One possible approach is to find a way to estimate either magnitude or distance independent of the attenuation relationships. Given either a magnitude or distance estimate, the constraints provided by the attenuation relationships, $(Y - \hat{Y}(M,R)) = 0$, can be used to find or constrain the other unknown. So, the question becomes, how can magnitude be estimated from observed P-wave amplitudes? Nakamura (1988) was the first to advocate use of the P-wave predominant period to estimate magnitude in seismic early warning applications. The central idea is that small and large earthquakes differ in the amount of low frequency energy they radiate; small earthquakes, which are like point sources, radiate relatively more high frequency energy than large events, which have more low frequency energy due to the finiteness of the rupture dimensions. Allen and Kanamori (2003) also use this idea of predominant period in estimating magnitude in their proposed early warning system for Southern California. In this thesis, a new method to estimate magnitude from ratios of ground motion envelope amplitudes is presented. Ratios of ground motion are also indicative of frequency content. In the frequency domain, displacement and acceleration are $X(\omega)$ and $\omega^2 X(\omega)$, respectively. The ratio of acceleration to displacement is

$$\frac{\ddot{X}(\omega)}{X(\omega)} = \frac{\omega^2 X(\omega)}{X(\omega)} = \omega^2 \tag{4.19}$$

I use linear discriminant analysis to find the relationship that optimally relates ground motions ratios to magnitude.

A short note on linear discriminant analysis

The following is a brief discussion of linear discriminant analysis, based on lectures from a course entitled "Methods in Applied Statistics and Data Analysis" taught by Emmanuel Candes and Tapio Schneider at Caltech.

Consider an $n \times m$ data matrix $\mathbf{X}$, where n corresponds to the number of earthquake records in the database, and m corresponds to the number of different types of observations, for instance $m = 3$, for log of acceleration, velocity, and displacement. (This is if we consider a single channel, say the vertical channel. This method is extendible to multi-channel analysis, for example, considering the vertical and both horizontal channels.) In linear discriminant analysis, the goal is to find a linear combination $\mathbf{X} \cdot u$ for which the separation between predefined groups is maximized. In the single-channel case (take the vertical channel), then u is an unknown 3×1 vector that is determined by maximizing the ratio of among-group variance to within-group variance. That is, we seek u that maximally separates the data of different groups while minimizing the spread within a given group. (In the seismic early warning application, the groups will be defined in terms of magnitude. The question being addressed is: what linear combination of the data $\mathbf{X} \cdot u$ best separates small and large earthquakes? An in-depth analysis of the seismic early warning application follows the presentation of the theory.)

Suppose there are g groups $G_i, i = 1 \ldots g$ with n_i observations in each group. Let $n = \sum\limits_i n_i$ be the total number of observations. Let $g(j)$ indicate the group to which the j^{th} observation belongs.

The *group means*, μ_i, are the means of the different types of observations within a group. They are defined as

$$\mu_i = \frac{1}{n_i} \sum_{j \in G_i} \mathbf{X}_{j:} \tag{4.20}$$

The notation $j:$ is used to denote the j^{th} row. So, $\mathbf{X}_{j:}$ is a row vector corresponding to the j^{th} row of $\mathbf{X}$. For each i, μ_i is a $1 \times m$ row vector corresponding to row mean of the observations within the i^{th} group.

The *within group covariance matrix*, $\mathbf{S^{(i)}}$, is defined as

$$\mathbf{S^{(i)}} = \frac{1}{n_i - 1} \sum_{j \in G_i} \left(\mathbf{X}_{j:} - \mu_i\right)^T \left(\mathbf{X}_{j:} - \mu_i\right) \tag{4.21}$$

The *pooled within group covariance matrix*, $\mathbf{S_w}$, is a weighted sum of within group covariance matrices, and is defined as

$$\mathbf{S_w} = \frac{1}{n - g} \sum_i (n_i - 1) \mathbf{S^{(i)}} \tag{4.22}$$

The *among group covariance matrix*, $\mathbf{S_a}$, is defined as

$$\begin{aligned}
\mathbf{S_a} &= \frac{1}{g-1}\sum_{i=1}^{g} n_i \left(\mu_i - \mu\right)^T \left(\mu_i - \mu\right) \\
\text{with the } \textit{grand mean}, \quad \mu & \quad \text{defined as} \\
\mu &= \frac{1}{n}\sum_{j=1}^{n} \mathbf{X}_{j:} = \frac{1}{n}\sum_{i} n_i \mu_i \qquad (4.23)
\end{aligned}$$

(Note: the *grand mean* is the mean of the different types of observations. Like the group means, it is also a $1 \times m$ row vector.)

The goal in linear discriminant analysis is to find a linear combination $\mathbf{X} \cdot u$ of the data such that the different groups are maximally separated, while observations within a group are maximally clustered. That is, we want to find the vector u that maximizes a separability measure, λ, which is the ratio of *among group* over *within group* variance.

$$\lambda = \frac{u^T \mathbf{S_a} u}{u^T \mathbf{S_w} u} = \frac{\text{among group variance}}{\text{within group variance}} \qquad (4.24)$$

The maximum of λ satisfies the condition $\frac{\partial \lambda}{\partial u} = 0$

$$\begin{aligned}
\frac{\partial \lambda}{\partial u} &= \frac{u^T \mathbf{S_a} u \cdot u^T \mathbf{S_w}}{\left(u^T \mathbf{S_w} u\right)^2} - \frac{u^T \mathbf{S_w} u \cdot u^T \mathbf{S_a}}{\left(u^T \mathbf{S_w} u\right)^2} = 0 \\
&= \lambda \cdot \frac{u^T \mathbf{S_w}}{u^T \mathbf{S_w} u} - \frac{u^T \mathbf{S_a}}{u^T \mathbf{S_w} u} = 0 \\
&\Rightarrow \lambda \cdot u^T \mathbf{S_w} - u^T \mathbf{S_a} = 0 \qquad (4.25) \\
&\quad \text{taking the transpose of Eqn. 4.25} \\
&\quad \text{and using the symmetry of covariance matrices } \mathbf{S_a} \text{ and } \mathbf{S_w} \\
&\quad \lambda \cdot \mathbf{S_w} u - \mathbf{S_a} u = 0 \\
&\Rightarrow \mathbf{S_a} u = \lambda \cdot \mathbf{S_w} u \\
&\quad \text{assuming } \mathbf{S_w} \text{ is invertible} \\
&\Rightarrow \mathbf{S_w^{-1}} \cdot \mathbf{S_a} u = \lambda \cdot u \qquad (4.26)
\end{aligned}$$

Eqn. 4.26 is an eigenvalue problem. The weight vector u is an eigenvector of

$\mathbf{S_w^{-1}} \cdot \mathbf{S_a}$; the separability measure λ is the eigenvalue of u.

Applying linear discriminant analysis to P-wave amplitudes

Linear discriminant analysis can be used to constrain magnitude estimates from available P-wave observations. (This is part of the effort to define the Bayesian likelihood function, $prob(Y|M, R)$.) The various quantities in linear discriminant analysis will be defined in terms of the quantities involved in seismic early warning. In particular, we are interested in estimating magnitude from ratios of different components of ground motion. Therefore, the columns of the data matrix $\mathbf{X}$ should correspond to the different components of ground motion. Since we want to maximize the available warning time, we want our estimates to be based on the P-wave signal rather than the S-wave. Thus, the columns of the data matrix $\mathbf{X}$ are defined to be the log of the P-wave envelope amplitudes fit to the envelope histories in the database. In this particular example, the columns correspond to log acceleration, log velocity, and log displacement of the vertical channel. The vertical channel is used since P-waves usually have larger amplitudes in the vertical direction. There are a total of 3373 vertical records in the envelope database. Thus n, the total number of observations, is 3373. $\mathbf{X}$ is thus a 3373×3 matrix. (If we are interested in including the horizontal channels, extra columns corresponding to the additional channels can be appended to the data matrix $\mathbf{X}$. Thus, if we wish to include the root mean square of the horizontal channels, $\mathbf{X}$ will be a 3373×6 matrix. If we want to include the North-South and East-West channels separately, $\mathbf{X}$ will be a 3373×9 matrix.)

Since we are interested in estimating magnitude, the groups are defined according to magnitude: $M < 3$, $3 \leq M < 4$, $4 \leq M < 5$, $5 \leq M < 6$, and $M \geq 6$.

With these group definitions and the above definition of the data or observation matrix $\mathbf{X}$, the procedure prescribed by Eqns. 4.20 through 4.26 is followed. The

Group number G_i	Magnitude range	Number of records n_i
1	$M < 3$	72
2	$3 \leq M < 4$	1094
3	$4 \leq M < 5$	1439
4	$5 \leq M < 6$	623
5	$M \geq 6$	145

Table 4.1: Group definitions for linear discriminant analysis of vertical P-wave amplitudes as magnitude indicators.

eigenvalues and eigenvectors of $\mathbf{S_w^{-1}} \cdot \mathbf{S_a}$ are:

$$\begin{aligned} \lambda_1 = 2674.2 \quad , \qquad & u_1^T = \begin{bmatrix} -0.23 & -0.2 & 0.95 \end{bmatrix} \\ \lambda_2 = 46.2 \quad , \qquad & u_2^T = \begin{bmatrix} -0.88 & 0.23 & 0.41 \end{bmatrix} \\ \lambda_3 = 9.9 \quad , \qquad & u_3^T = \begin{bmatrix} -0.41 & 0.82 & -0.4 \end{bmatrix} \end{aligned} \tag{4.27}$$

Figure 4.1 shows the eigenvectors of $\mathbf{S_w^{-1}} \cdot \mathbf{S_a}$. Since $\lambda_1 = 2674.2$ is by far the largest eigenvalue, the first eigenvector $u_1^T = \begin{bmatrix} -0.23 & -0.2 & 0.95 \end{bmatrix}$ gives the linear combination of the vertical ground motion amplitudes that is optimally indicative of magnitude. Figure 4.2 shows the linear combinations of the data $Z_1 = \mathbf{X} \cdot u_1$ and $Z_2 = \mathbf{X} \cdot u_2$ corresponding to the first two eigenvectors of $\mathbf{S_w^{-1}} \cdot \mathbf{S_a}$. From this Figure, $Z_1 = \mathbf{X} \cdot u_1$ gives a better separation of the various magnitude groups. There is less overlap of the different groups in the Z_1 axis than along the Z_2 axis. From this application of linear discriminant analysis, the linear combination $Z_1 = \mathbf{X} \cdot u_1 = -0.23 \log(acc) - 0.2 \log(vel) + 0.95 \log(disp)$ is that which maximizes the ratio of among group to within group variances, or which optimally separates the different magnitude groups while minimizing the spread within each group. (Recall that the columns of $\mathbf{X}$ correspond to $\log(acc), \log(vel), \log(disp)$.)

Figure 4.3 summarizes the linear discriminant analysis of P-wave amplitudes as indicators of magnitude using $\log(acc), \log(vel), \log(disp)$. (These results use acceleration, velocity, and displacement; shortly, results using just the acceleration and

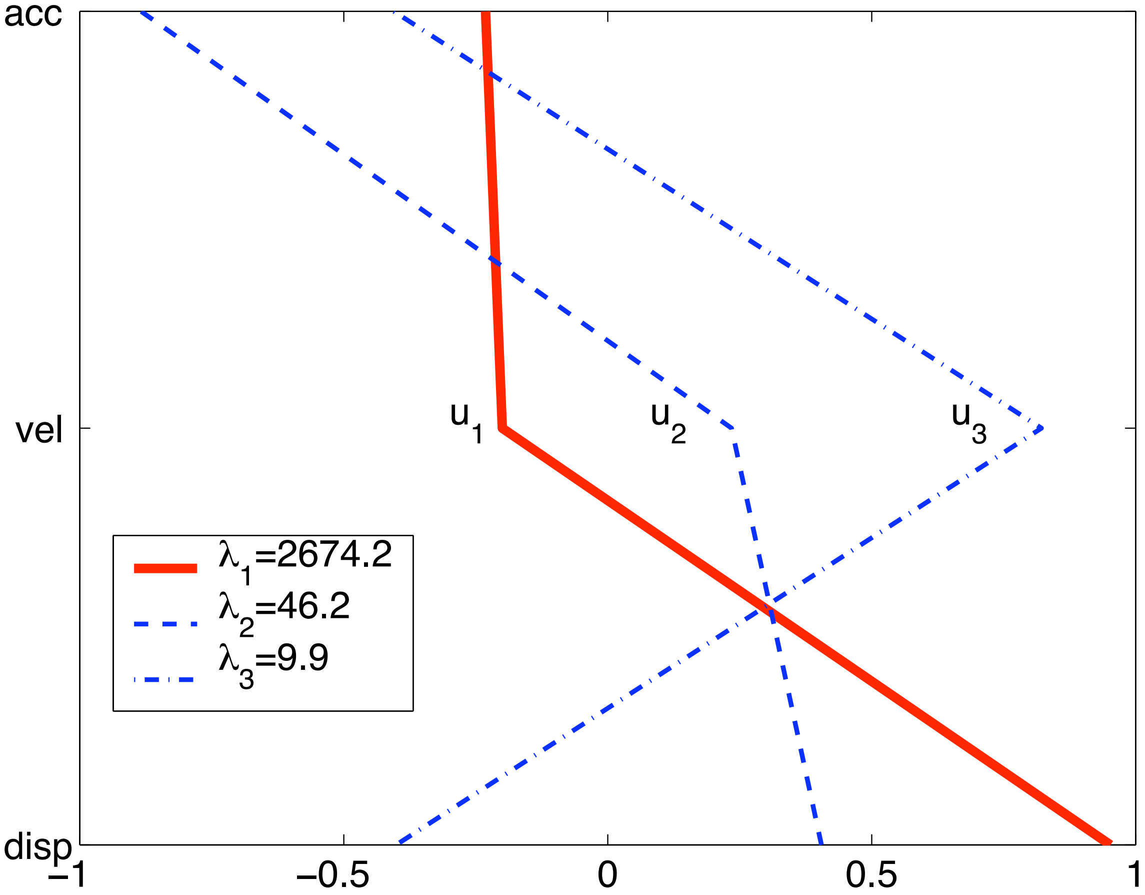


Figure 4.1: Eigenvectors of $\mathbf{S_w^{-1}} \cdot \mathbf{S_a}$. The y-axis corresponds to $\log(acc), \log(vel), \log(disp)$; the x-axis are the coefficients obtained via the linear discriminant analysis. Since $\lambda_1 = 2674.3$ is by far the largest eigenvalue, its eigenvector u_1 provides the linear combination $\mathbf{X}u_1$ that best separates the different magnitude groups while maximizing the clustering within each magnitude group.

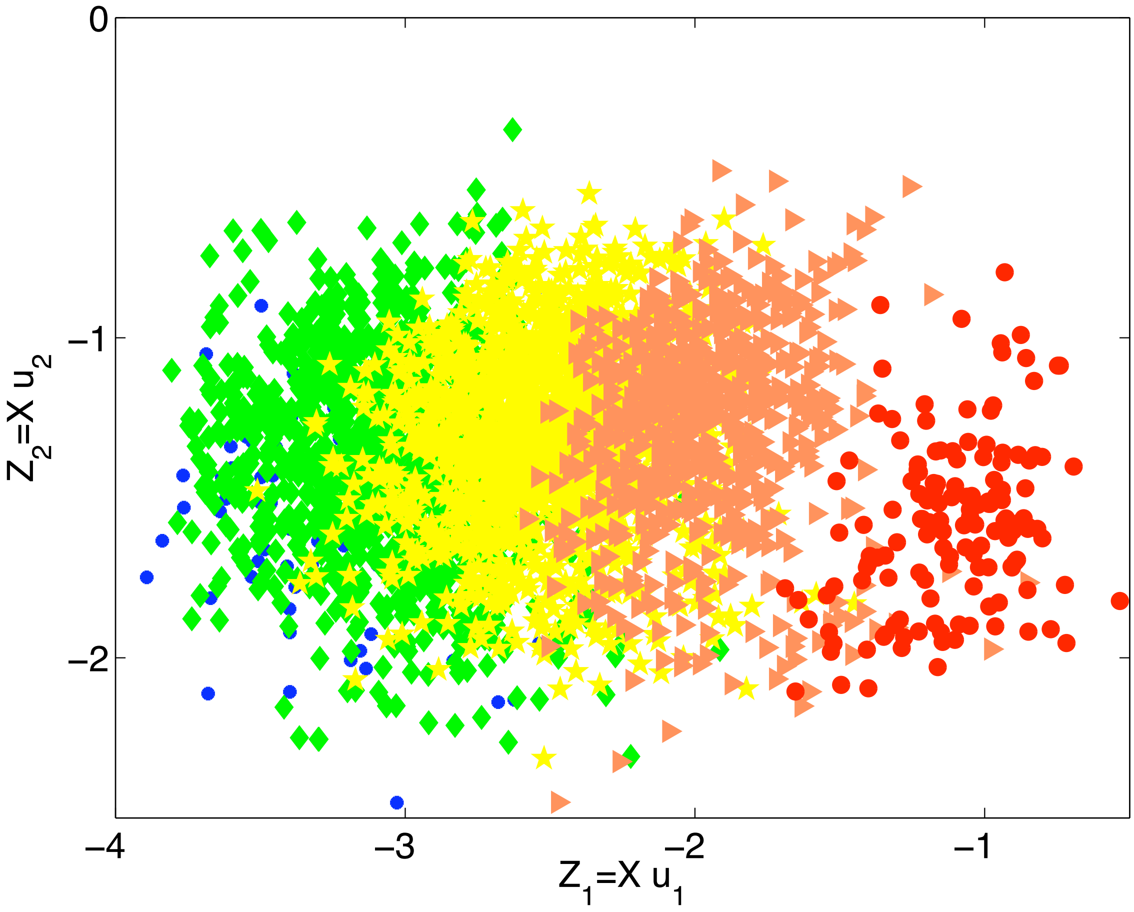


Figure 4.2: Linear combinations of the data $\mathbf{X}$ corresponding to the first and second eigenvectors of $\mathbf{S_w^{-1}} \cdot \mathbf{S_a}$. $Z_1 = \mathbf{X} \cdot u_1$ has better separation of the various magnitude groups than $Z_2 = \mathbf{X} \cdot u_2$.

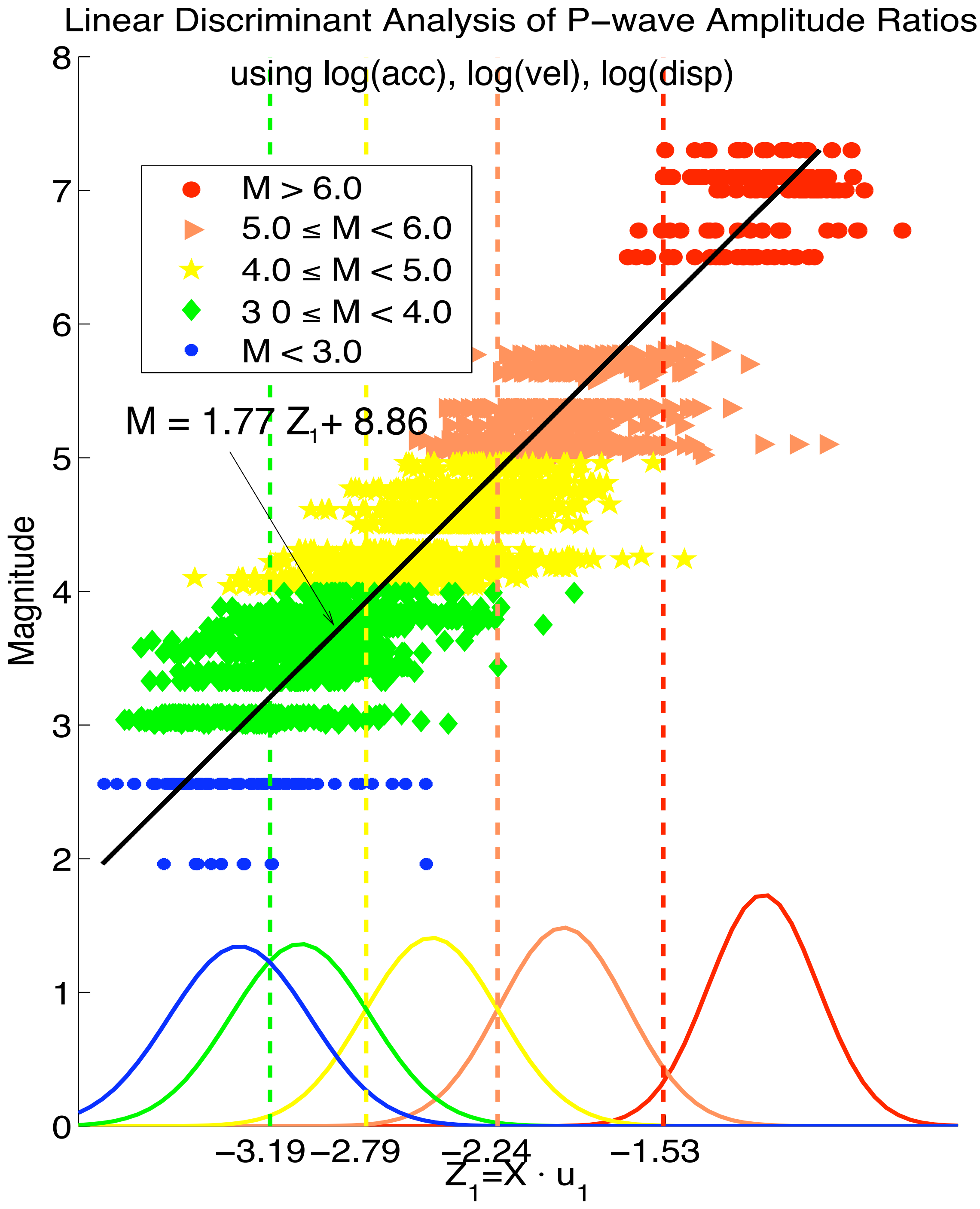


Figure 4.3: Linear discriminant analysis on P-wave amplitude ratios (using $\log(acc), \log(vel), \log(disp)$).

displacement observations will be presented.) On the x-axis is the optimal linear combination, $Z_1 = \mathbf{X} \cdot u_1 = -0.23 \log(acc) - 0.2 \log(vel) + 0.95 \log(disp)$; "true" magnitude (known from SCSN catalogue) is on the y-axis. The Normal curves at the bottom of the plot are the best-fit normal curves to the projection of the observations in each group onto the eigenvector u_1. The dashed lines are called *decision boundaries.* They are located at the midpoint between the means of the projections onto u_1 of two adjacent groups. The decision boundaries from this analysis are $Z_{1-2} = -3.19$ between groups 1 and 2, $Z_{2-3} = -2.79$ between groups 2 and 3, $Z_{3-4} = -2.24$ between groups 3 and 4, and $Z_{4-5} = -1.54$ between groups 4 and 5. Decision boundaries are useful in classifying new observations. New observations are classified according to which group mean (after projecting onto u_1) they are closest to. For example, let $\mathbf{X}_{new}$ be a vector of new observations (log acceleration, velocity, and displacement), and $Z_{new} = \mathbf{X}_{new} \cdot u_1$ be the projection of these new observations onto u_1. Based on Z_{new}, the new event is in group 1 ($M < 3$) if $Z_{new} < Z_{1-2}$; it is in group 2 ($3 \leq M < 4$) if $Z_{1-2} \leq Z_{new} < Z_{2-3}$, in group 3 ($4 \leq M < 5$) if $Z_{2-3} \leq Z_{new} < Z_{3-4}$, etc.

Table 4.2 is the *confusion matrix* for the linear discriminant analysis of P-wave amplitudes as indicators of magnitude. The rows denote the actual group; the columns denote the group into which the data would be classified based on the LDA results. Each cell has two numbers. The top number is the percentage of the time that a particular classification is made for data in a given group. The bottom number in parentheses are the number of observations classified in a particular group (denoted by the column number). The values on the diagonal denote how often the correct classification is made. The off-diagonal entries describe the misclassification error. For example, for events in the magnitude range $4 \leq M < 5$ (group 3), the decision boundaries from the LDA analysis yield the correct classification 68% of the time. There is a 15% and 16% chance of misclassifying an event in this magnitude range as either a group 2 ($3 \leq M < 4$) or a group 4 ($5 \leq M < 6$) event, respectively. There is about a 1% chance that it will be classified as $M \leq 3$ or $M \geq 6$. The last column is the number of observations in each group.

If a strict interpretation of linear discriminant analysis results is followed, given

Confusion matrix for LDA of P-wave amplitudes as magnitude indicators using acceleration, velocity, and displacement
Row: actual group; column: classification based on LDA

	Group 1	Group 2	Group 3	Group 4	Group 5	Total obs.
Group 1	70% (50)	24% (17)	7% (5)	0% (0)	0% (0)	=72
Group 2	33% (356)	51% (559)	16% (175)	$< 1\%$ (4)	0 (0)	=1094
Group 3	1% (13)	15% (218)	68% (978)	16% (229)	$< 1\%$ (1)	=1439
Group 4	0% (0)	0% (0)	15% (94)	78% (488)	7% (41)	=623
Group 5	0% (0)	0% (0)	0% (0)	3% (5)	97% (140)	=145

Table 4.2: Confusion matrix for linear discriminant analysis of vertical P-wave amplitudes as magnitude indicators (using acceleration, velocity, and displacement). The confusion matrix provides an idea of how often the decision boundaries from the linear discriminant analysis correctly classify (or misclassify) the data. Recall that group 1 is $M < 3$, group 2 is $3 \leq M < 4$, group 3 is $4 \leq M < 5$, group 4 is $5 \leq M < 6$, and group 5 is $M \geq 6$.

a new set of observations, only the group to which an event most likely belongs can be determined. For example, it can be said that a new event has magnitude between $5 \leq M < 6$, or that $M > 6$, based on where on the Z_1 axis the new observations fall.

However, from Figure 4.3, there is a strong correlation between $Z_1 = \mathbf{X} \cdot u_1$ on the x-axis and magnitude on the y-axis. A best fit linear relationship between Z_1, the optimal linear combination of ground motion amplitudes, and M can be found. This is the solid black line going through the data points in Figure 4.3. It is given by

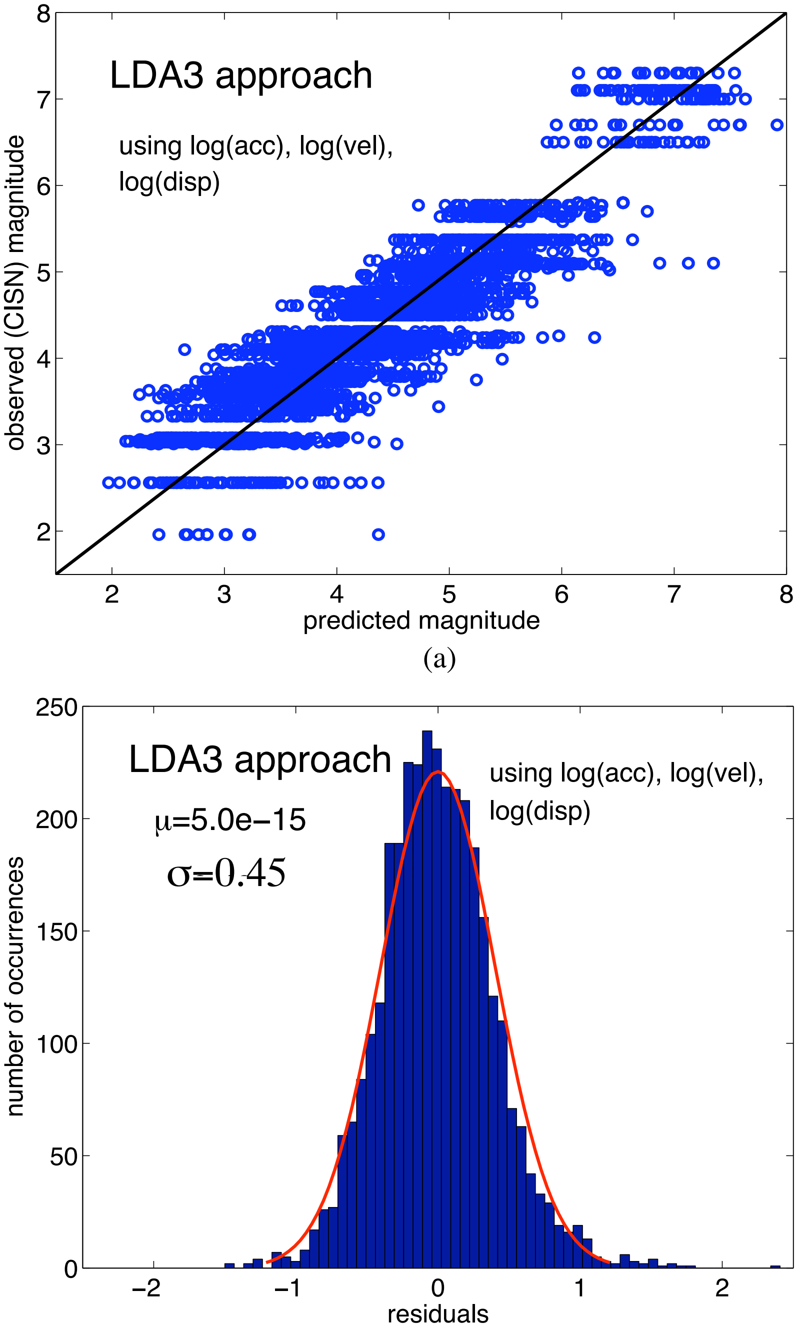
$$\begin{aligned} \hat{M}_{LDA_3} &= 1.77 Z_1 + 8.86 \\ &= 1.77\left(-0.23\log(acc) - 0.2\log(vel) + 0.95\log(disp)\right) + 7.98 \\ &= -0.41\log(acc) - 0.35\log(vel) + 1.68\log(disp) + 8.86 \qquad (4.28) \end{aligned}$$

The advantage of using Eqn. 4.28 is that given the available ground motion amplitudes after a station has triggered (which are presumably P-waves), a magnitude estimate, $\hat{M}_{LDA_3}$, can immediately be obtained. (With a strict interpretation of LDA results, we can only make statements about which magnitude group the new event is most likely to belong to.) The standard error of the regression is given by $\hat{\sigma}_{LDA_3}$, where

$$\hat{\sigma}^2_{LDA_3} = \frac{\sum_{j=1}^{n} \left(M_i - \hat{M}_{LDA_3,i} \right)^2}{ndof} \qquad (4.29)$$

$ndof$ is $n - 2 = 3373 - 2 = 3371$, since we are solving for a slope (coefficient on Z_1) and an intercept. $\hat{\sigma}_{LDA_3}$ is 0.45 magnitude units. The residuals are defined as $M_i - \hat{M}_{LDA_3,i}$.

Figure 4.4 shows some regression diagnostics for Eqn. 4.28. In Figure 4.4(a), the magnitudes reported by SCSN for the earthquakes in the database are plotted against those predicted by Eqn. 4.28. Recall that Eqn. 4.28 related magnitude to P-wave acceleration, velocity, and displacement. The solid black line is the true regression line. It has a slope of 1. If Eqn. 4.28 predicted magnitudes exactly, then all the blue circles would fall on the solid black line. Thus, deviations of the blue circles from the true regression line are indicative of the prediction errors of Eqn. 4.28.



(b)

Figure 4.4: Regression diagnostics for Eqn. 4.28, which relates magnitude to P-wave acceleration, velocity, and displacement. In subplot (a), the observed or "true" magnitudes are plotted against those predicted by Eqn. 4.28. The straight line is the true regression line. In subplot (b), the residuals $(M_i - \hat{M}_{LDA_3,i}$ between the "true" SCSN magnitudes and those predicted by Eqn. 4.28 are shown; they are normally distributed. The mean is practically 0, the variance is 0.45.

Figure 4.4(b) shows histograms of the residuals $M_i - \hat{M}_{LDA_3,i}$ and the best-fit Normal curve to the residuals. From this plot, the residuals are normally distributed. We can thus speak of confidence intervals. The 95% confidence interval of Eqn. 4.28 is $\hat{M}_{LDA_3} \pm 2\hat{\sigma}_{LDA_3}$. That is, given observations of vertical P-wave acceleration, velocity, and displacement and the method described above, the interval $\hat{M}_{LDA_3} \pm 2\hat{\sigma}_{LDA_3}$ or $\hat{M}_{LDA_3} \pm 0.9$ contains the correct magnitude of the earthquake 95% of the time. This uncertainty is quite large, nearly 1 magnitude unit. However, recall that this is from just P-wave observations at a single station. As more data becomes available, either additional P-wave data at the given station, the S-wave arrival, or arrivals at other stations, the uncertainty on the magnitude and location estimates will improve. How additional data enters the estimation process and reduces the uncertainties on the source estimates is an important issue that will be addressed shortly.

An alternative approach: simple linear regression

In Eqn. 4.28, we performed a regression for the optimal linear combination of the data matrix, $Z_1 = \mathbf{X}u_1$ on magnitude, and found a relationship that relates magnitude to the log of P-wave acceleration, velocity, and displacement. Alternatively, a model can be postulated relating magnitude directly to log of P-wave amplitudes, such as

$$M_j = \alpha \log(acc_j) + \beta \log(vel_j) + \gamma \log(disp_j) + \delta + \epsilon_j \qquad (4.30)$$

where $\alpha, \beta, \gamma, \delta$ are regression coefficients that can be determined via simple linear regression, and ϵ is statistical or prediction error, which we assume to be normally distributed with mean 0 and constant variance σ. (This is a standard assumption on the statistical errors in regression problems. This assumption was used extensively in the regressions in Chapter 2.) The subscript j is used to be consistent with the notation used in the linear discriminant analysis.

In matrix notation, Eqn. 4.30 is

$$\begin{bmatrix} M_1 \\ \vdots \\ M_j \\ \vdots \\ M_n \end{bmatrix} = \begin{bmatrix} \log(acc_1) & \log(vel_1) & \log(disp_1) & 1 \\ \vdots & \vdots & \vdots & 1 \\ \log(acc_j) & \log(vel_j) & \log(disp_j) & 1 \\ \vdots & \vdots & \vdots & 1 \\ \log(acc_n) & \log(vel_n) & \log(disp_n) & 1 \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \begin{bmatrix} \epsilon_i \\ \vdots \\ \epsilon_j \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{4.31}$$

or

$$\mathbf{M} = \mathbf{X_1} \cdot \boldsymbol{\alpha} + \boldsymbol{\epsilon} = \hat{\mathbf{M}} + \boldsymbol{\epsilon} \tag{4.32}$$

where $\boldsymbol{\alpha}^T = \begin{bmatrix} \alpha & \beta & \gamma & \delta \end{bmatrix}$. Note that the matrix $\mathbf{X_1}$ is simply the matrix $\mathbf{X}$ in the linear discriminant analysis with a column of 1's appended. The least squares estimates of $\boldsymbol{\alpha}$ are given by

$$\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{X}_1^T \cdot \boldsymbol{X}_1\right)^{-1} \cdot \boldsymbol{X}_1^T \cdot \mathbf{M} \tag{4.33}$$

The covariance matrix of $\hat{\boldsymbol{\alpha}}$ is given by

$$\begin{aligned} cov(\hat{\boldsymbol{\alpha}}) &= \hat{\sigma}^2 \left(\mathbf{X_1}^T \cdot \mathbf{X_1}\right)^{-1} \qquad (4.34) \\ \text{where} \quad & \hat{\sigma}^2 = \frac{RSS}{\text{degrees of freedom}} = \frac{RSS}{n-p} \\ \text{and} \quad & RSS = \sum_{j=1}^{n} \left(\hat{M}_j - M_j\right) \\ & p = \text{ number of unknown coefficients} = 4 \\ & n = \text{ total observations in database} = 3373 \end{aligned}$$

The variances of the unknowns, $\hat{\boldsymbol{\sigma}}^2_{\boldsymbol{\alpha}} = \begin{bmatrix} \hat{\sigma}^2_\alpha & \hat{\sigma}^2_\beta & \hat{\sigma}^2_\gamma & \hat{\sigma}^2_\delta \end{bmatrix}^T$, correspond to the diagonal entries of the covariance matrix $cov(\hat{\boldsymbol{\alpha}})$. The standard errors (or standard deviations) are the square roots of the variances.

Table 4.3 shows the least squares estimates of $\boldsymbol{\alpha}^T = \begin{bmatrix} \alpha & \beta & \gamma & \delta \end{bmatrix}$, the standard errors (square root of variance) on each of the coefficients, the lower and upper bounds of the approximate 95% confidence intervals, and the P-values. The 95% confidence interval is an interval that contains the true value of the parameter being estimated 95% of the time. For normally distributed data, this is typically given by $\pm 2\sigma$. The P-value expresses the probability that, due to random errors, a given coefficient takes a non-zero value when it is in fact 0. In statistics terminology, it is the probability of rejecting the null hypothesis when the null hypothesis is in fact true. P-values close to 0 indicate that a given predictor is statistically significant. Larger P-values mean that there is a relatively larger probability that a given parameter is equal to zero even though the regression analysis yields a non-zero value. Table 4.3 shows that only $\hat{\beta}$ (the coefficient for $\log(vel)$) has a non-zero P-value. This means that the most significant predictors of magnitude are $\log(acc)$, $\log(disp)$, and a constant term. There is a 3% chance of our regression analysis showing that the coefficient for magnitude dependence on $\log(vel)$ is $\hat{\beta} = -0.16$ when in fact β may be zero. Despite the $\log(vel)$ term having a non-zero P-value, all predictors in Eqn. 4.32 are significant at the 5% level.

Least squares estimates for simple linear regression
of magnitude on $\log(acc)$, $\log(vel)$, and $\log(disp)$

	estimate	standard error	lower bound	upper bound	P-value
$\hat{\alpha}$	-0.42	0.04	-0.5	-0.34	0
$\hat{\beta}$	-0.16	0.08	-0.36	0	0.03
$\hat{\gamma}$	1.32	0.04	1.24	1.40	0
$\hat{\delta}$	8.05	0.05	7.95	8.15	0

Table 4.3: Least squares estimates for unknowns in Eqn. 4.33, which postulates a linear relationship between P-wave log amplitudes and magnitude.

Thus, direct regression of magnitude onto the P-wave log amplitudes (acceleration, velocity, and displacement) yields

$$\hat{M}_{SLR,3} = -0.42\log(acc) - 0.16\log(vel) + 1.32\log(disp) + 8.05 \tag{4.35}$$

with SLR denoting “simple linear regression”. The relationship obtained via linear discriminant analysis is given by Eqn. 4.28, repeated here

$$\hat{M}_{LDA} = -0.41\log(acc) - 0.35\log(vel) + 1.68\log(disp) + 8.86$$

From Table 4.3, there is log acceleration and log velocity coefficients from linear discriminant analysis (Eqn. 4.28) are within the 95% confidence intervals of those from the simple regression. There are discrepancies in the log displacement coefficient and constant. Residual diagnostics for the simple linear regression results are shown in Figure **??**. The linear regression relationship (Eqn. 4.35) consistently underpredicts $M > 6$ events. For this reason, the relationship based on linear discriminant analysis (Eqn. 4.28) is preferred.

Best estimates of $\hat{M}, \hat{R}$ given P-wave acceleration and displacement

A similar type of analysis is performed using just P-wave acceleration and displacement, since from Table 4.3, these are the most statistically significant terms. In this case, the data matrix $\mathbf{X}_2$ is a 3373×2 matrix, with $\log(acc)$ and $\log(disp)$ as its two columns. The linear combination of P-wave $\log(acc)$ and $\log(disp)$ that is optimally indicative of magnitude is given by the eigenvector $\mathbf{u}$ corresponding to the largest eigenvalue γ of Eqn. 4.26

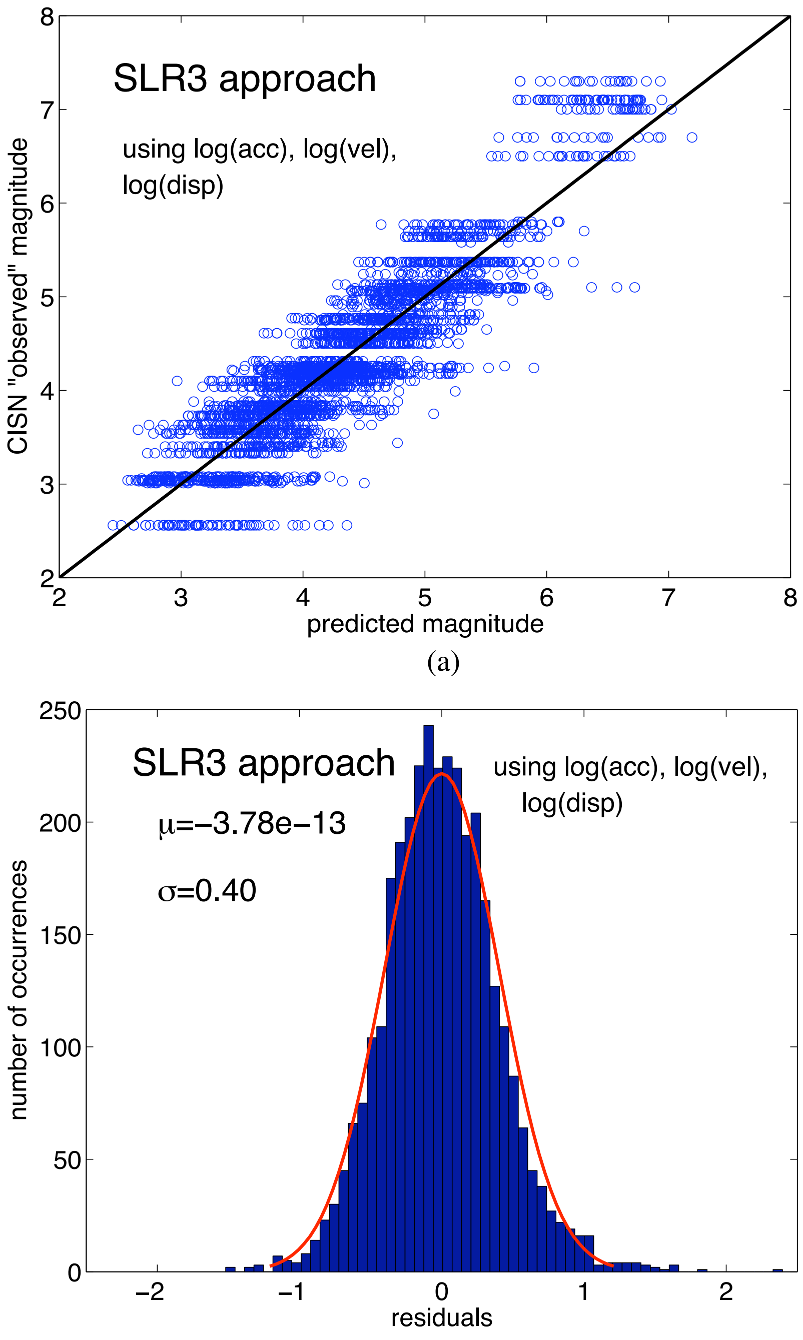
$$\mathbf{S_w^{-1}} \cdot \mathbf{S_a} u = \lambda \cdot u$$

With $\mathbf{X}_2$ being a 3373×2 matrix, it has 2 eigenvectors.

$$\begin{aligned} \lambda_1 = 2663.1 \quad , \qquad & u_1^T = \begin{bmatrix} 0.36 & -0.93 \end{bmatrix} \\ \lambda_2 = 46 \quad , \qquad & u_2^T = \begin{bmatrix} -0.83 & 0.56 \end{bmatrix} \end{aligned} \tag{4.36}$$

The confusion matrix for the linear discriminant analysis using only $\log(acc)$ and $\log(disp)$ is fairly similar to the confusion matrix for the analysis considering acceleration, velocity, and displacement.

The linear combination of $\log(acc)$ and $\log(disp)$ that is optimally indicative of magnitude is $Z = X_2 \cdot u_1 = 0.36\log(acc) - 0.93\log(disp) = acc^{0.36}/disp^{0.93}$. A



(a)

(b)

Figure 4.5: Regression diagnostics for Eqn. 4.35, which regresses magnitude directly onto P-wave acceleration, velocity, and displacement. In subplot (a), the observed or “true” magnitudes are plotted against those predicted by Eqn. 4.35. Systematic deviations from the true regression line (solid like with slope=1) give an idea of which magnitude ranges this relationship is applicable. Eqn. 4.35 systematically underpredicts events $M \geq 6$, and systematically overpredicts $M \leq 3$. Also similar to Eqn. 4.28, the residuals between the “true” SCSN magnitudes and those predicted by Eqn. 4.35 are normally distributed. The mean is practically 0, the standard deviation is 0.4.

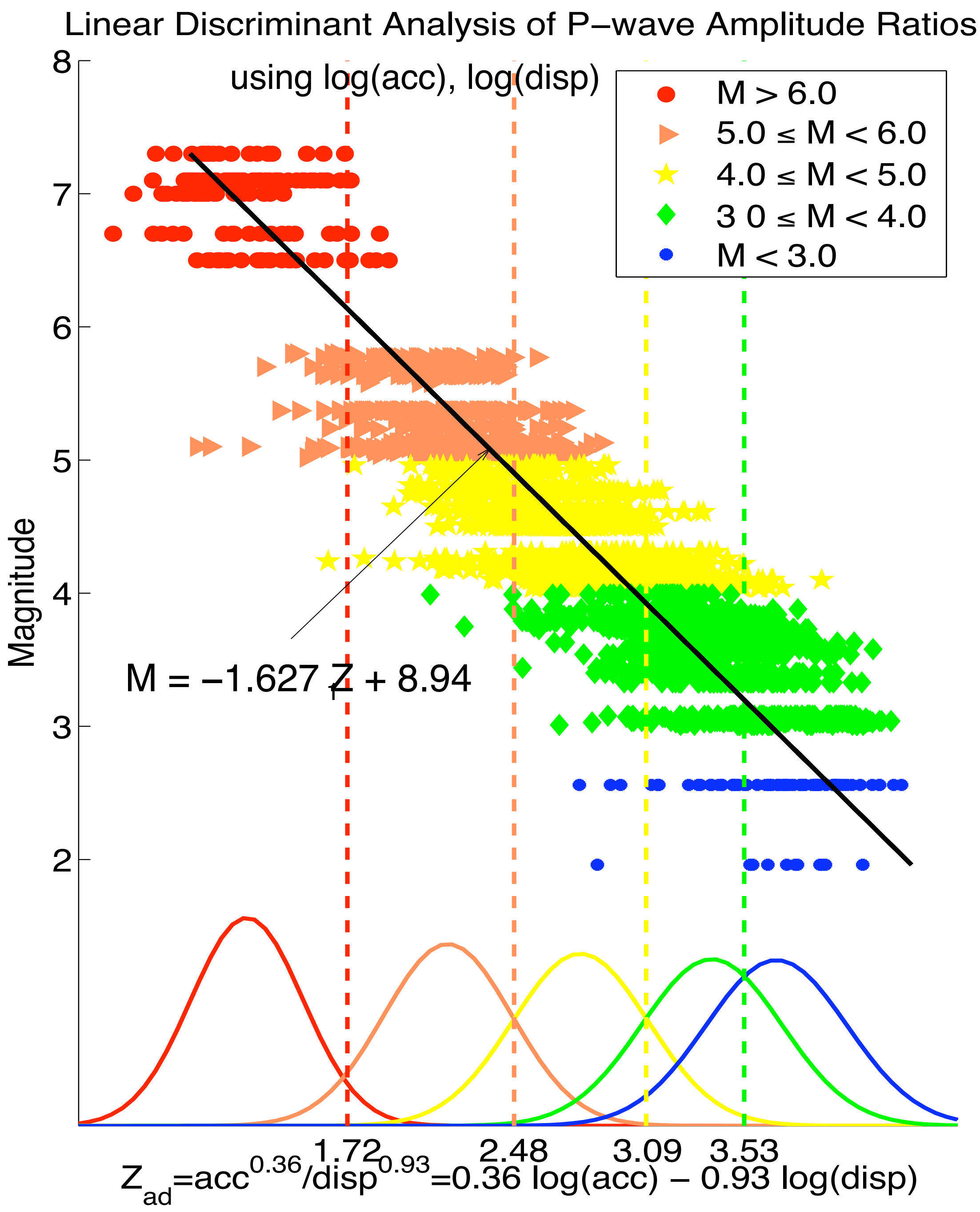


Figure 4.6: Linear discriminant analysis of P-wave $\log(acc)$ and $\log(disp)$ as indicators of magnitude. $Z = X_2 \cdot u = 0.36\log(acc) - 0.93\log(disp)$.

Confusion matrix for LDA of P-wave amplitudes as magnitude indicators
using acceleration and displacement
Row: actual group; column: classification based on LDA

	Group 1	Group 2	Group 3	Group 4	Group 5	Total obs.
Group 1	71% (51)	24% (17)	6% (4)	0% (0)	0% (0)	=72
Group 2	32% (345)	53% (575)	16% (170)	< 1% (4)	0 (0)	=1094
Group 3	1% (12)	15% (215)	68% (985)	16% (226)	< 1% (1)	=1439
Group 4	0% (0)	0% (0)	15% (93)	79% (492)	6% (38)	=623
Group 5	0% (0)	0% (0)	0% (0)	6% (8)	94% (137)	=145

Table 4.4: Confusion matrix for linear discriminant analysis of vertical P-wave $\log(acc)$ and $\log(disp)$ as magnitude indicators. Recall that group 1 is $M < 3$, group 2 is $3 \leq M < 4$, group 3 is $4 \leq M < 5$, group 4 is $5 \leq M < 6$, and group 5 is $M \geq 6$.

relationship between magnitude and $Z = acc^{0.36}/disp^{0.93}$ can be obtained:

$$\begin{aligned}\hat{M}_{LDA2} &= -1.627Z + 8.94 \\ &= -1.627(0.36\log(acc) - 0.93\log(disp)) + 8.95 \\ &= -0.59\log(acc) + 1.51\log(disp) + 8.94 \qquad (4.37)\end{aligned}$$

The standard error of regression on Eqn. 4.37 is $\sigma_{LDA2} = 0.45$.

Least squares estimates for simple linear regression
of magnitude on $\log(acc)$ and $\log(disp)$

	estimate	standard error	lower bound	upper bound	P-value
$\hat{\alpha}$	-0.50	0.02	-0.54	-0.46	0
$\hat{\gamma}$	1.24	0.02	1.20	1.28	0
$\hat{\delta}$	8.09	0.04	8.01	8.17	0

Table 4.5: Confusion matrix for linear discriminant analysis of vertical P-wave $\log(acc)$ and $\log(disp)$ as magnitude indicators. Recall that group 1 is $M < 3$, group 2 is $3 \leq M < 4$, group 3 is $4 \leq M < 5$, group 4 is $5 \leq M < 6$, and group 5 is $M \geq 6$.

This linear discriminant analysis (LDA) and regression methods presented in this Section are based on P-wave amplitudes. The relationships obtained here are applicable to P-waves. The results will be in error if the input amplitudes are actually S-wave amplitudes. Similar analyses were performed on S-wave amplitudes. The details are in Appendix F. Interestingly, the S-wave ground motion ratio that is optimally indicative of magnitude is identical to the P-wave ground motion ratio. Thus, the ratio $Z_{ad} = 0.36log(acc) - 0.93log(disp) = acc^{0.36}/disp^{0.93}$ is optimally indicative of magnitude, whether the amplitudes are from P- or S-waves. However, there are differences in the relationships that relate the ground motion ratio Z_{ad} to magnitude, depending on whether the amplitudes are from P- or S-waves.

$$\begin{aligned} \hat{M}_P &= -1.627Z_{ad} + 8.94, \sigma_{M_P} = 0.45 \text{ for P-wave amplitudes,} && (4.38) \\ \hat{M}_S &= -1.459Z_{ad} + 8.05, \sigma_{M_S} = 0.41 \text{ for S-wave amplitudes,} && (4.39) \\ \text{where } Z_{ad} &= 0.36log(acc) - 0.93log(disp) \\ &= acc^{0.36}/disp^{0.93} \end{aligned}$$

It is therefore important to be able to distinguish between P- and S-waves. A method (also based on linear discriminant analysis) developed to distinguish between P- and S-waves is presented in Appendix E.

Comment 1: A robust magnitude estimation approach would be to assume that the amplitudes are always from S-waves. Using the S-wave relationships on P-wave amplitudes will initially underestimate the magnitude. However, this is corrected once the S-wave arrives. The magnitude estimates will consistently approach the actual magnitude from below. It is possible that such an approach will decrease the occurrence of false alarms.

Comment 2: In retrospect, the ratio of peak vertical acceleration to peak horizontal displacement may perhaps be a better ratio to use than just the vertical acceleration to displacement. Using the combined vertical and horizontal ratio might perform better that that using just the vertical channel once the S-wave arrives. This approach does not require distinguishing between P- and S-waves.

Thus far, two ways to use the initially observed waveform data to estimate magnitude and distance have been presented. In this Section, the use of 1) envelope attenuation relationships to quantify the trade-offs between magnitude and location and 2) linear discriminant analysis and simple linear regression to estimate magnitude from available amplitudes were discussed. These analyses used only the vertical channel. Extension of the analysis to include the horizontal channels is straightforward.

Initial form of the likelihood function, $prob(Y|M,R)$

The likelihood function, $prob(Y|M,R)$, given the first available observations of acceleration, velocity, and displacement from a given channel will be defined using the envelope attenuation relationships and the linear discriminant analysis methods. The following assumptions are required: 1) ground motion amplitudes are log-normally distributed; 2) acceleration, velocity, and displacement envelope amplitudes are independent quantities; 3) amplitudes observed at time t_1 are independent of those observed at a later time t_j; 4) amplitudes observed at station A are independent of those at station B. The first assumption is valid, as was shown by the regression diagnostics for the envelope amplitude attenuation relationships in Chapter 2. The other assumptions are on more tenuous ground. It can be argued that while the observed envelope amplitudes for different channels, times, and stations are not *causatively* independent (since they are caused by the same earthquake), they can be considered *stochastically* independent (Beck, J. personal communication). That is, knowing the peak acceleration at a given time does not exactly determine peak velocity or displacement. Knowledge of one quantity does not imply knowledge of the others. (On the other hand, if an event has large peak accelerations, the other amplitudes are likely to be large as well.) These assumptions are necessary to proceed.

Let $Y_{A,t_1}^T = \begin{bmatrix} Y_a & Y_v & Y_d \end{bmatrix}$ be the initial set of log P-wave acceleration, velocity, and displacement observed at time t_1 at station A. The envelope attenuation relationships and linear discriminant analysis magnitude estimators will be used to find the magnitudes and locations most consistent with these observations Y_{A,t_1}. Recall that Bayes' theorem states that

$$prob(M,R|Y_{A,t_1}) \propto prob(Y_{A,t_1}|M,R) \times prob(M,R) \tag{4.40}$$

To focus attention on the likelihood function, $prob(Y_{A,t_1}|M,R)$, assume for now a uniform prior density function over the range of possible magnitudes and distances (or locations), or $prob(M,R) = constant$. The most probable estimates of M,R given the available observations Y_{A,t_1} are those that maximize the posterior density function, $prob(M,R|Y_{A,t_1})$. (The t_1 subscript will be dropped for the moment.) Given a uniform prior, these M,R are those that maximize the likelihood function, $prob(Y_A|M,R)$. Eqn. 4.40 becomes

$$\begin{aligned} prob(M,R|Y_A) &\propto prob(Y_A|M,R) \\ &= prob(Y_a,Y_v,Y_d|M,R) \\ &= prob(Y_a|M,R)\cdot prob(Y_v|M,R)\cdot prob(Y_d|M,R) \end{aligned} \tag{4.41}$$

Since linear discriminant analysis (LDA) showed that acceleration and displacement ratios are best indicative of magnitude, the acceleration and displacement amplitudes, Y_a, Y_d will be included into the likelihood function via the LDA magnitude estimators and the velocity amplitude Y_v via the envelope attenuation relationships for velocity.

The likelihood of observing a P-wave velocity amplitude Y_v is given by

$$prob(Y_v|M,R) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_v}\exp\left(-\frac{(Y_v - \bar{Y}_v(M,R))^2}{2\hat{\sigma}_v^2}\right) \tag{4.42}$$

where Y_v is the observed P-wave velocity, and $\bar{Y_v}(M,R)$ is the P-wave amplitude envelope attenuation relationship (similar to Eqn. 2.3, with the P-wave velocity coefficients). Acceleration and displacement amplitudes Y_a, Y_d are included via the linear discriminant analysis relationships.

$$\begin{aligned} prob(Y_a, Y_d|M,R) &\propto prob(Z_{ad}|M) \qquad (4.43) \\ &\propto \frac{1}{\sqrt{2\pi}\hat{\sigma}_{LDA_2}} \exp\left(-\frac{(Z_{ad}-\hat{Z}_{ad}(M))^2}{2\hat{\sigma}^2_{LDA_2}}\right) \qquad (4.44) \end{aligned}$$

Eqn. 4.44 holds because the groups in the linear discriminant analysis were defined solely in terms of magnitude. The LDA, as set up, does not yield any information about location, though ground motion ratios have a statistically significant distance dependence. Eqn. 4.44 uses a rearranged form of the LDA relationships; it models the ground motion ratios as normally distributed. Figure 4.8 shows that the residuals $\hat{M}_{LDA_2} - M_{SCSN}$ are normally distributed with 0 mean and constant variance.

Therefore, the likelihood of observing $Y^T_{A,t_1} = \begin{bmatrix} Y_a & Y_v & Y_d \end{bmatrix}$ is given by

$$\begin{aligned} prob(Y_A|M,R) &= prob(Y_a, Y_v, Y_d|M,R) \\ &\propto prob(Y_a, Y_d|M,R) \cdot prob(Y_v|M,R) \\ &\propto \frac{1}{2\pi\hat{\sigma}_{LDA2}\hat{\sigma}_v} \exp\left(-\frac{2\hat{\sigma}^2_{LDA2}(Y_v - \bar{Y}_v(M,R))^2 + 2\hat{\sigma}^2_v(M-\hat{M}_{LDA2})^2}{4\hat{\sigma}^2_{LDA2}\hat{\sigma}^2_v}\right) \qquad (4.45) \\ &\propto prob(M,R|Y_A) \qquad (4.46) \end{aligned}$$

Taking no prior information into account, the M, R that maximize $prob(Y_A|M,R)$ are also those that maximize the posterior density function, $prob(M,R|Y_A)$, and therefore are the most probable source estimates M, R given the initially available observed P-wave data from station A, $Y^T_{A,t_1} = \begin{bmatrix} Y_a & Y_v & Y_d \end{bmatrix}$. Thus, the problem of estimating M, R from the initial observations $Y^T_{A,t_1} = \begin{bmatrix} Y_a & Y_v & Y_d \end{bmatrix}$ from a single station reduces to a problem of minimizing the term

$$2\hat{\sigma}^2_{LDA2}(Y_v - \bar{Y}_v(M,R))^2 + 2\hat{\sigma}^2_v(M-\hat{M}_{LDA2})^2 \qquad (4.47)$$

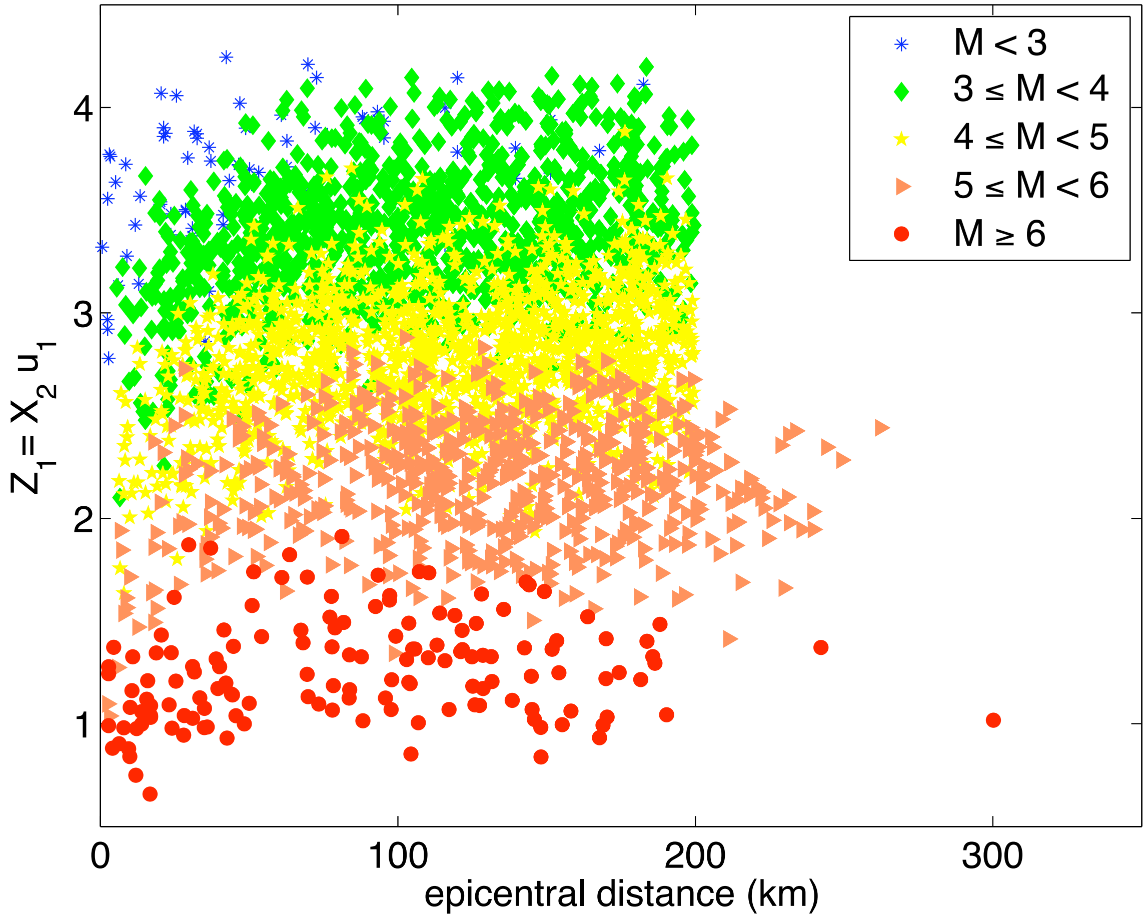
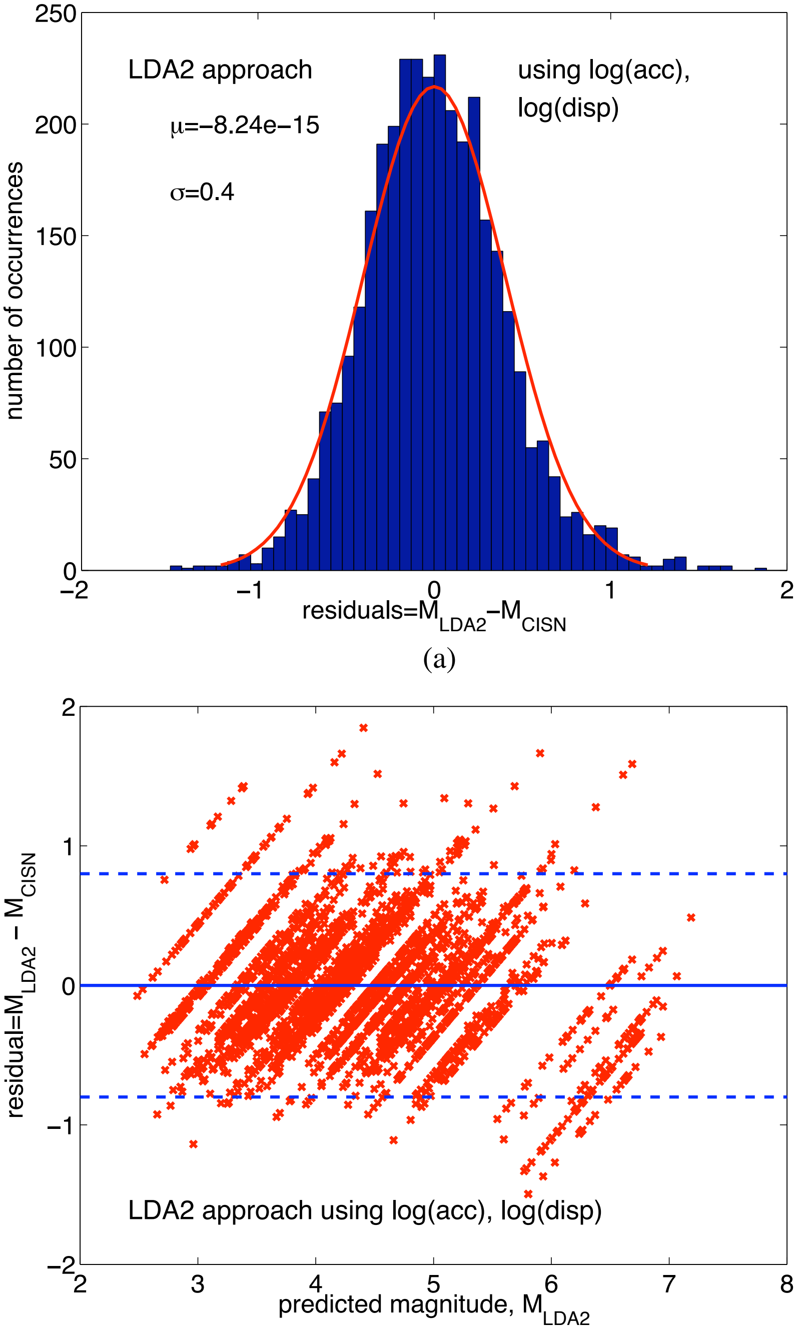


Figure 4.7: The ground motion ratio, Z_1, has a slight distance dependence.



(a)

(b)

Figure 4.8: Some regression diagnostics of $\hat{M}_{LDA_2}$. These justify modeling the ground motion ratios Z_{ad} as normally distributed.

Inclusion of additional data

Thus far, we have discussed how to estimate magnitude and epicentral distance from a set of peak P-wave acceleration, velocity, and displacement amplitudes from a single channel (vertical) on a single station. In particular, the M, R that maximize the right hand side of Eqn. 4.45 (or minimize Eqn. 4.47) are the most probable source estimates given the first available envelope data at time t_1 from station A, $Y_{A,t_1}^T = \begin{bmatrix} Y_a & Y_v & Y_d \end{bmatrix}$. In Eqn. 4.45, the observed velocity amplitude Y_v and the P-wave velocity attenuation relationship are used to quantify the trade-offs between magnitude and distance. Using just the attenuation relationships leave these trade-offs unresolved, since the constraint $(Y_v - \bar{Y}_v(M, R)) = 0$ is satisfied by an infinite set of M, R pairs. However, the linear discriminant analysis for magnitude (or magnitude-based group) limits the magnitude range that is consistent with the observed ground motion ratio, $Z_{ad} = acc^{0.36}/disp^{0.93}$.

Incorporating additional information will now be addressed. As time elapses, more waveform data will become available as 1) the slower-traveling phases arrive at the first triggered station and as 2) ground motions propagate to and trigger stations further from the source region. This additional data contains the information necessary to resolve the magnitude-distance trade-offs. For example, as the S-wave arrives at the first triggered station, a distance estimate from the waveform data can be obtained from the S-P time. In the beginning of the estimation process, the available waveform data alone may not be able to resolve the trade-offs between magnitude and location. It is in this situation (when the set of available waveform data is sparsely populated) that prior information is most useful. The role of prior information and the different types of prior information relevant to seismic early warning will be discussed in the following Section. Including later arriving data will help resolve the trade-offs in the initial source estimates, and increase the reliability (or reduce the variances) on these estimates. The following questions will be addressed: 1) how to use all available envelope data, not just the peak amplitudes, 2) how to include data from other stations, and 3) how to relate the estimates from the data available at time t_1 to the estimates from the data available at some later time, t_2?

Including all available envelope data

Let $Y_{A,t_1:t_n}$ be a $n \times 1$ vector representing all the envelope data of a given channel (say, vertical acceleration) available at station A, in the time interval $t_1 \leq t \leq t_n$, (assuming that the sampling rate is 1 sample per second). Bayes' theorem states that

$$prob(M, R|Y_{A,t_1:t_n}) \propto prob(Y_{A,t_1:t_n}|M, R) \times prob(M, R) \tag{4.48}$$

If we are to use Eqn. 4.48, we need to make a slight modification to our ground motion model from Chapter 2, Eqn. 2.1. Let us explicitly account for the statistical (or prediction error) and say that

$$E_{observed}(t) = E(t, M, R) + \epsilon \tag{4.49}$$

where

$$E(t, M, R) = \sqrt{E_P^2(t) + E_S^2(t) + E_{ambient}^2} \tag{4.50}$$

Let $Y_{A,t_1:t_n} = \begin{bmatrix} Y_{A,t_1} & \cdots & Y_{A,t_n} \end{bmatrix} = E_{observed}(t)$. Assume that the statistical errors ϵ in Eqn. 4.49 are normally and independently distributed with zero mean and constant variance σ_E^2, and that the set of ground motion amplitudes observed at a given time t_i are independent from those at some different time t_k. $prob(Y_{A,t_1:t_n}|M, R)$ an be written as

$$\begin{aligned} prob(Y_{A,t_1:t_n}|M, R) &\propto \prod_{k=1}^{n} prob(Y_{A,t_k})|M, R) \qquad (4.51)\\ &\propto \frac{1}{(\sqrt{2\pi})^n \hat{\sigma}_E^n} \exp\left(-\sum_{k=1}^{n}\left(\frac{(Y_{A,t_k} - E(t_k, M, R))^2}{2\hat{\sigma}_E^2}\right)\right) \qquad (4.52)\end{aligned}$$

Assuming a uniform prior, the M, R pair that is most probable given the available observed envelope data for a given channel at station A in the time interval $t_1 \leq t \leq t_n$,

$Y_{A,t_1:t_n}$ is the pair that minimizes the quantity

$$\sum_{k=1}^{n}(Y_{A,t_k} - E(t_k, M, R))^2 \tag{4.53}$$

It may be of interest to find the M, R pair that is most probable given the observed envelope data from *all available channels*. Assuming that the different channels are independent (again, one of the tenuous assumptions), the only modification required to Eqn. 4.53 is the addition of another index (let us use j) corresponding to summing over all available channels:

$$\sum_{j=1}^{m}\left(\sum_{k=1}^{n}(Y_{A,t_k} - E(t_k, M, R))^2\right)_j \tag{4.54}$$

where m is the number of available channels.

How useful is this? The initial estimates from maximizing Eqn. 4.45 (or minimizing Eqn. 4.47) given the first available P-wave data at station A are available, the envelopes of later-arriving phases on the various channels can be predicted. Once these slower phases arrive at station A, the initial estimates can be adjusted by minimizing either Eqn. 4.53 or 4.54, depending on the number of channels included. This method can be easily extended to include waveform data from stations further from the source region. These relationships provide the basis for real-time Kikuchi-Kanamori type inversions using envelopes of ground motion.

Including data from additional stations

If the source estimation problem is phrased in terms of magnitude and epicentral distances, having data from p stations means there are $p + 1$ unknowns to solve for: the magnitude estimate $\hat{M}$, and p epicentral distances of the earthquake source to each of p stations, $R_1, \cdots, R_p$. When there are more than two stations, it is beneficial to reparameterize the estimation problem in terms of magnitude, latitude, and longitude of the earthquake. Any occurrence of epicentral distance terms R_p are replaced with

$$R_p = K\sqrt{(lat_p - lat_e)^2 + [(lon_p - lon_e)\cos\frac{lon_p + lon_e}{2}]^2} \tag{4.55}$$

where $K = 111.13$ km/degree, lat_p, lon_p are latitude and longitude in degrees of the p^{th} station, lat_e, lon_e are the estimated latitude and longitude of the earthquake. Assuming that the data from the various stations are independent, data from other stations can be included by appending multiplicative terms to Eqn. 4.51. This is equivalent to adding an additional index corresponding to summing over the available stations to Eqn. 4.53. The M, lat_e, lon_e that minimize

$$\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{l=1}^{p}(Y_{l,t_k} - E_l(t_k, M, lat_e, lon_e))_j^2 \tag{4.56}$$

where n is the number of seconds since the P-wave arrival at the first triggered station, m is the number of available channels (i.e., vertical and/or horizontal acceleration, velocity, and displacement), and p is the number of stations included in the estimation process. The M, lat_e, lon_e that minimize Eqn. 4.56 are those those that maximize the likelihood function. With a uniform prior, these source estimates M, lat_e, lon_e also maximize the posterior density function, and are thus the most probable source estimates given all the available data.

The advantages of using a geographic coordinate system are 1) there are 3 unknowns no matter how many stations are included in the estimation process (as opposed to $p+1$ unknowns involved in keeping track of epicentral distance estimates for p stations) and 2) the prior information relevant to seismic early warning is most efficiently dealt with in a geographic coordinate system.

Updating estimates as additional data becomes available

The third question raised at the beginning of this section was: how are estimates updated as additional data becomes available? One way to proceed would be to minimize Eqn. 4.56 at each time a new estimate is desired, with the index k going from 1 (corresponding to the time of the P-wave trigger at the first station) to n (corresponding to the time at which the current estimate would be made). However, this procedure can become unwieldy as the amount of data available increases. Bayes' theorem provides a more efficient approach to updating estimates because its calculations are sequential (Sivia, 1996).

For the case of an observation available at station A at time t_1, Y_{A,t_1}, Bayes' theorem states that

$$prob(M, R|Y_{A,t_1}) \propto prob(Y_{A,t_1}|M, R) \times prob(M, R) \tag{4.57}$$

Assuming a uniform prior over the magnitude and distance range in question,

$$prob(M, R|Y_{A,t_1}) \propto prob(Y_{A,t_1}|M, R) \tag{4.58}$$

When a new observation is available at time $t_2 > t_1$, the posterior pdf taking into account the observations at t_1 and t_2 is

$$prob(M, R|Y_{A,t_1}, Y_{A,t_2}) \propto prob(Y_{A,t_1}, Y_{A,t_2}|M, R) \tag{4.59}$$

Assuming that the observations at t_1 and t_2 are independent (tenuous), then

$$prob(Y_{A,t_1}, Y_{A,t_2}|M, R) \propto prob(Y_{A,t_2}|M, R) \times prob(Y_{A,t_1}|M, R) \tag{4.60}$$

Eqn. 4.59 then becomes

$$\begin{aligned} prob(M, R|Y_{A,t_1}, Y_{A,t_2}) &\propto prob(Y_{A,t_1}, Y_{A,t_2}|M, R) \\ &\propto prob(Y_{A,t_2}|M, R) \times prob(Y_{A,t_1}|M, R) \\ &\quad \text{using Eqn. 4.58} \\ &\propto prob(Y_{A,t_2}|M, R) \times prob(M, R|Y_{A,t_1}) \end{aligned} \tag{4.61}$$

Eqn. 4.61 is in the form *posterior* $\propto$ *likelihood* $\times$ *prior*, with $prob(M, R|Y_{A,t_1})$

as the prior pdf. However, $prob(M, R|Y_{A,t_1})$ is also the posterior pdf accounting for only the observation at time t_1. Thus, the posterior pdf taking into account the observations at times t_1 and t_2 is given by the likelihood of the observation at t_2 times the posterior taking into account only the observation at t_1. This is important with regards to the question of updating estimates. This means that if we desire a new estimate at time t_{n+k}, where we have k seconds worth of data since the previous estimate at time t_n, rather than having to minimize Eqn. 4.54 with the time index going from 1 to $n+k$, to update the estimate, we need only to minimize over the time index $n+1$ to $n+k$ and use the estimate at t_n as the prior. This is more efficient, since it takes advantage of the previously performed calculations. It is a nice characteristic of Bayesian calculations that sequential processing of the incoming data provides the same answer as a simultaneous analysis of all the data (Sivia, 1996).

4.3.1 A summary of the likelihood function, $prob(Y|M, R)$

In the subsequent Chapters, the likelihood will be defined in terms of the ratio between vertical acceleration and displacement amplitudes and the envelope attenuation relationships for vertical acceleration and horizontal acceleration, velocity, and displacement. Maximizing the likelihood function will be equivalent to minimizing the term $L(M, lat, lon)$:

$$\begin{align}
L(M, lat, lon) &= \sum_{i=1}^{n} \sum_{j=1}^{\text{P,S}} L(M, lat, lon)_{ij} \tag{4.62} \\
L(M, lat, lon)_{ij} &= \frac{(Z_{ad_{ij}} - \bar{Z}_j(M))^2}{2\sigma^2_{Z_{ad_{ij}}}} + \sum_{k=1}^{4} \left(\frac{(Y_{obs_{ijk}} - \bar{Y}_{ijk}(M, lat, lon))^2}{2\sigma^2_{ijk}} \right) \tag{4.63} \\
i &= 1 \ldots n, \text{ where n is the number of stations with P detections} \nonumber \\
j &= 1, 2 \text{ ,for body wave phases (P- and S-waves)} \nonumber \\
k &= 1, \ldots 4, \text{ for peak amplitudes of vertical velocity, and} \nonumber \\
& \quad \text{horizontal acceleration, velocity, and displacement} \nonumber \\
Z_{ad_{ij}} &= 0.36 \log(PZA_{ij}) - 0.93 \log(PZD_{ij}) \nonumber \\
PZA_{ij} &= \log_{10} \text{ of peak vertical acceleration for phase j at station i} \nonumber \\
PZD_{ij} &= \log_{10} \text{ of peak vertical displacement for phase j phase at station i} \nonumber \\
Y_{obs_{ijk}} &= \log_{10} \text{ of peak observed amplitude} \nonumber \\
& \quad \text{of k channels and phase j at station i} \nonumber \\
\bar{Z}_j(M) &= \alpha_j M + \beta_j \tag{4.64} \\
R_i^2 &= \text{epicentral distance between an earthquake located at } lon, lat \nonumber \\
& \quad \text{and station i} \nonumber \\
R_{1_i} &= \sqrt{(R_i^2 + 9)} \nonumber \\
C(M)_{jk} &= c_{1_{jk}}(\arctan(M - 5) + 1.4) \times \exp c_{2_{jk}}(M - 5) \nonumber \\
\bar{Y}_{ijk}(M, R) &= a_{jk}M - b_{jk}(R_{1_i} + C(M)_{jk}) - d_{jk} \log_{10}(R_{1_i} + C(M)_{jk}) + e_{ijk} \tag{4.65}
\end{align}$$

The assumptions regarding independence were necessary to equate maximizing the likelihood function with minimizing Eqn. 4.62. Eqn. 4.62 holds for multiple stations. A single station estimate involves setting $n = 1$, and replacing epicentral location lat, lon with a single epicentral distance, R. A minimum of 3 seconds of data since the P arrival and 2 seconds of data since the S arrival at a given station is required before that station contributes P- and S-wave amplitudes to the source estimation process. It is assumed that P-waves can be detected via short-term over long-term average methods. The best source estimates M, lat, lon at any given time are those most probable given the peak P- and S-wave amplitudes available at that time. P- and S-waves can be distinguished via the P/S discriminants described in Appendix F.

When sufficient amplitude observations are available, the likelihood function has a global maximum. The M, lat, lon that maximize the likelihood function (or minimize Eqn. 4.62) are those which the ground motion models (LDA magnitude estimators and envelope attenuation relationships) indicate are consistent with the available amplitude observations. These M, lat, lon are comparable to Kanamori (1993)'s strong motion centroid. The location estimate is an amplitude-based location. As discussed by Kanamori (1993), such amplitude-based locations are more robust than arrival-based locations and are an efficient means to convey the spatial distribution of ground motion for post-earthquake response. When the set of available observations is sparse (for instance, 3 seconds after the initial P detection at the first triggered station), the likelihood function may not have a global maximum; there may be trade-offs between the source estimates that are unresolved by the available observations. It is such situations that the Bayes prior is useful.

4.4 Defining the prior, $prob(M, R)$

Different types of information can be included in the Bayes prior, $prob(M, R)$. In the seismic early warning problem, the prior is information related to relative earthquake probabilities that may aid in the source estimation problem. It is a statement regarding our best knowledge (in this case, of earthquake occurrence) about the problem before examining the waveform data from the on-going rupture.

Seismological considerations

A uniform prior implies that earthquakes of all magnitudes are equally likely at all distances or locations. While a uniform prior simplifies the calculations involved, it is not an accurate statement about relative earthquake probabilities.

- Many earthquakes occur on known faults (though new faults are still being discovered).

- Earthquakes cluster in time and space. For example, large earthquakes are typically followed by aftershocks in the source region, with Omori's law governing the decay of number of aftershocks as a function of time since the mainshock. Omori's law is an empirical relationship that states

  $$n = \frac{C}{(K+t)^P} \tag{4.66}$$

  where n is frequency of aftershocks at time t after the mainshock. K, C, P are constants determined for a particular mainshock (Lay and Wallace, 1995).

- The Gutenberg-Richter law generally governs the magnitude-frequency distribution of earthquakes.

  $$\begin{aligned} \log N(M) &= A - bM \\ N &= 10^A \times 10^{-bM} \end{aligned} \tag{4.67}$$

  where $N(M)$ is the number of earthquakes with magnitudes in the interval $(M, M+\Delta M)$, A is the *activity* and is related to the maximum magnitude possible, and the b is the *b-value*, which is typically around 1. The Gutenberg-Richter relationship states that there are 10 times more earthquakes of magnitude M than earthquakes of magnitude $M+1$.

Reasenberg and Jones (1989) combine Omori's law and the Gutenberg-Richter relationship to express the probability of one or more events within the magnitude range ($M_1 \leq M < M_2$) and within the time range ($S \leq t < T$) after the mainshock as

$$P = 1 - \exp\left[- \int_{M_1}^{M_2} \int_{S}^{T} \lambda(t, M)\, dt\, dM \right] \tag{4.68}$$

$$\text{where} \qquad \lambda(t, M) = 10^{a+b(M_n - M)}(t+c)^{-p} \tag{4.69}$$

where $\lambda(t, M)$ is the rate of aftershocks with magnitude M or greater at time t after a mainshock of magnitude M_n.

- Many large earthquake have foreshocks. In a study of 59 $M \geq 5$ earthquakes, Abercrombie and Mori (1996) found that 44% of the earthquakes in their California dataset had foreshocks; they also found that foreshock occurrence is a function of mainshock rake and depth, but not of mainshock magnitude. The findings of Abercrombie and Mori are consistent with those of an earlier study by Jones (1984). Jones studied 20 mainshocks in the San Andreas system and found that 35% were preceded by foreshocks in the immediate spatial and temporal vicinity (within 1 day, 5 km of the mainshock).

In addition to these factors, geometric considerations and the state of health of the seismic network monitoring the earthquake activity can provide useful constraints to the seismic early warning problem.

Geometric considerations

Some simple geometric considerations can be used to constrain the location estimates in seismic early warning. Assuming that earthquakes are equally likely to occur at all locations implies that they are more likely to be far away than in close. The probability of an event occurring at distance R is proportional to area of a ring or annulus with radius R and finite thickness dR, a reasonable approximation of which is the circumference of the circle times the thickness dR. From this consideration, the probability of a given epicentral distance is thus a linear function of distance.

$$Pr(R) = R \tag{4.70}$$

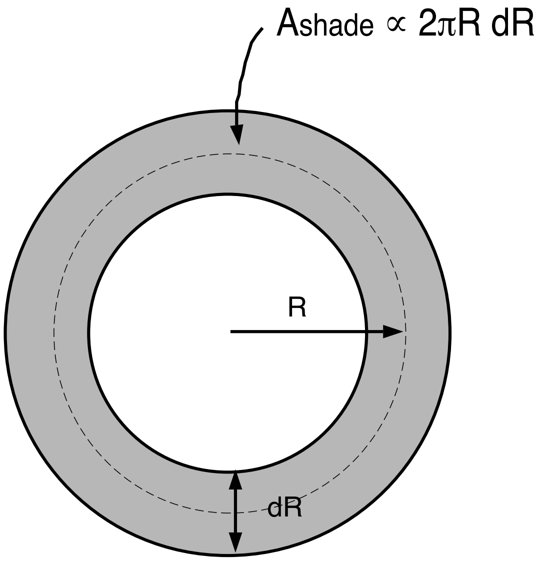


Figure 4.9: The probability of an earthquake (or any event) occurring at radius R is proportional to the circumference of the circle times a finite thickness dR.

The state of health of the seismic network also provides a geometric constraint to the location of an earthquake. In particular, Voronoi cells (Sambridge, 1999a) or nearest neighbor regions of the operating stations can be calculated. The Voronoi cell of a given station is the set of all location coordinates that are closer to that given station than any other station in the network. Figure 4.10 shows a set of 30 randomly generated points and their Voronoi cells. How are such Voronoi cells useful in seismic early warning? Voronoi cells should be calculated for the stations operating on any given day, and updated to reflect the current status of the various stations. Once an earthquake triggers a station, assuming that the Voronoi cells are an accurate snapshot of the operating stations within the network and P-waves can be detected accurately, the location is immediately constrained to be within the Voronoi cell of the first triggered station. The denser the seismic network, the smaller the area of the average Voronoi cell, and the stronger the constraint on location. The time elapsed between the first and subsequent arrivals at adjacent stations is also a useful constraint. If the ground motions trigger two stations simultaneously, then the earthquake location can be constrained to a line, the common edge shared by the Voronoi cells of the two stations. The longer the time elapsed between the P-wave arrival at the first station and the subsequent triggers at surrounding stations, the closer into the interior of the Voronoi cell of the first triggered station the earthquake must be located.

The use of not-yet-arrived data, as described by Rydelek and Pujol (2004) and Horiuchi et al. (2004), can be built upon to describe the evolution of the region of likely location with time after the first P detection. In this thesis, the region of likely location is the region consistent with the observed arrivals. It is independent of the location estimate obtained from maximizing the likelihood function; that location estimate is amplitude-based. From Rydelek and Pujol (2004), the locations consistent with the first two P arrivals (let us call them t_1, t_2) satisfy the equation $d_2 - d_1 = V(t_2 - t_1)$, where d_1, d_2 are the epicentral distances of the event to stations 1 and 2, and where V is an average P-wave velocity. The equation $d_2 - d_1 = V(t_2 - t_1)$ describes a hyperbola.

Consider the case where there is a P detection at station 1, and t_{est}, Δt seconds after the first P detection, there are still no P detections at the adjacent stations. Assume that the coordinates of operating stations are known, and that there are n stations sharing a Voronoi edge with station 1. For $i = 1, \ldots, n$, each of these n stations provides the constraint that

$$d_i - d_1 > V(t_{est} - t_1) = V(\Delta t) \tag{4.71}$$

The region of likely location is the intersection of the Voronoi cell of the first triggered station, and the areas consistent with n inequality constraints described by Eqn. 4.71. When $\Delta t = 0$ (the first P detection), this region corresponds to the Voronoi cell of the first triggered station. When $\Delta t > 0$ and there are no P-arrivals at the adjacent stations, the region of likely location is an area within the first station's Voronoi cell. How much smaller this area is than the original Voronoi cell is a measure of the power of constraints on location by Δt, and is thus a function of Δt. Once the P-wave arrives at the second station, this area collapses to Rydelek's (2004) hyperbola. A third arrival locates an epicenter.

Rydelek and Pujol (2004) and Horiuchi et al. (2004) use not-arrived data after the first two P arrivals. Given the Voronoi cells of the seismic network, not-yet-arrived data can be used to provide continuously evolving constraints on the region of likely location immediately after the first P detection. This is advantageous in regions with low station density, where the time between the first and second P arrivals may be relatively long. This will be illustrated more concretely in the subsequent Chapters.

The Voronoi cells involve prior information since they are based on station locations, which are known beforehand. The not-yet-arrived data is not strictly prior information, since the time Δt since the first P detection without subsequent arrivals is an observed quantity. However, it is not included in the likelihood since it does not involve observed amplitudes.

The use of these different types of prior information, seismological and geometric, and the not-yet-arrived data, will be illustrated in the examples in the following Chapters.

A balance between the prior and likelihood

Bayes' theorem is analogous to the human learning process. Before beginning an experiment, we may have certain beliefs or expectations. In Bayesian terms, these beliefs can be quantified by the Bayesian prior. Observations from the phenomena under study continually adjust these prior beliefs. If the incoming data is consistent with the prior beliefs, then it increases our degree of belief in our prior notions. However, if the incoming data exhibits inconsistencies with the prior beliefs, we make modifications to what we currently believe given the available data. As additional observations become available, our current beliefs are increasingly determined by the data, less so by the prior. This allows for the possibility that what is being observed

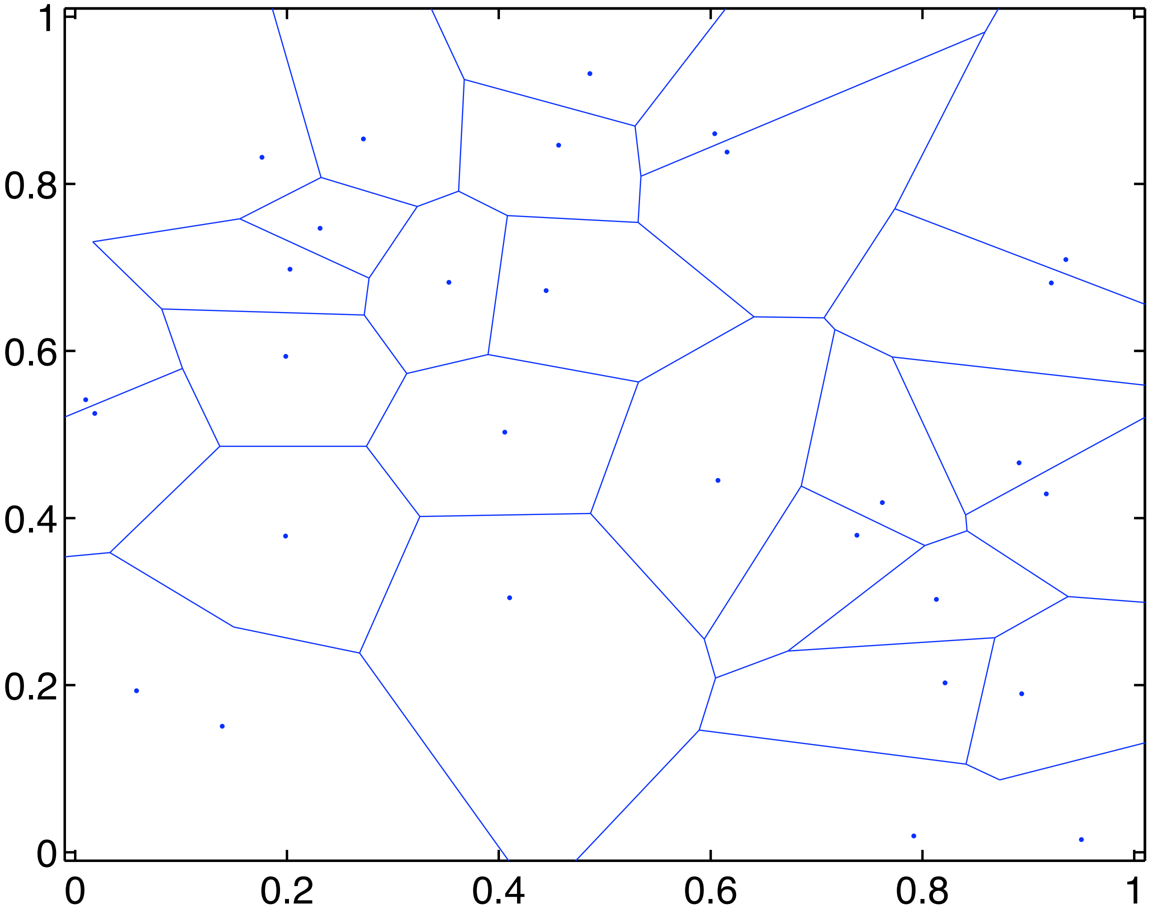


Figure 4.10: Voronoi cells of 30 randomly generated points.

is something new and unexpected. When all data has been gathered and analyzed, either of the two conclusions will be made: 1) the process in study behaved in accordance to expectations before the experiment, or 2) the results were not expected, and something new was learned.

This type of framework is well-suited to real-time applications, such as seismic early warning. Seismic early warning has somewhat contrasting objectives: *reliable* source estimates are desired as *quickly* as possible. To maximize the available warning time, it is necessary to make estimates from a very sparse set of observations (for example, from the first available P-wave amplitudes from the first triggered station). Given a sparse set of observations, there will be trade-offs among the source parameters that cannot be resolved by the observations alone. It is more rational to resolve these trade-offs to be consistent with the types of prior information relevant to seismic early warning, namely: previously observed seismicity, foreshock/aftershock statistics (or Gutenberg-Richter / Omori's law type relationships), and the state of health of the seismic network, than to not specify the prior. Not explicitly specifying a prior, or basing the estimation process on the likelihood function (or observed data) alone, is equivalent to stating that we believe that earthquakes of all magnitudes occur at all locations with equal probability, which is an inaccurate description of the general state of knowledge.

As more observations become available, the influence of the prior decreases and the estimates become more reliable. (This does not mean that using prior information makes estimates less reliable.) Once the observations are able to resolve the magnitude-location trade-off on their own (for example, with an S-wave arrival, or with triggers at multiple stations), the initial form of the prior becomes irrelevant.

Thus far, reliabilities have only been discussed in a qualitative manner. The ensuing (quantitative) discussion of reliabilities of the best estimates $\hat{M}, \hat{R}$ is based on Sivia (1996).

4.5 Reliabilities of best estimates $\hat{M}$, $\hat{R}$

Let us now address the question of variances on the estimates $\hat{M}$ and $\hat{R}$.

Sivia (1996) recommends working with $L = log_e(prob(M, R|Y))$. L is often called the "log-likelihood". Recall that the best estimates $\hat{M}, \hat{R}$ maximize $prob(M, R|Y)$ (as well as $L = log_e(prob(M, R|Y)))$, and satisfy Eqn. 4.72

$$\begin{aligned} \left.\frac{\partial L}{\partial M}\right|_{\hat{M},\hat{R}} &= 0 \qquad (4.72) \\ \left.\frac{\partial L}{\partial R}\right|_{\hat{M},\hat{R}} &= 0 \end{aligned}$$

As described by Sivia (1996), to examine the shape or spread of the posterior about the maxima, we take a Taylor series expansion of $L = log_e(prob(M, R|Y))$ about the maxima $(\hat{M}, \hat{R})$.

$$\begin{aligned} L = L(\hat{M}, \hat{R}) + \frac{1}{2}\Bigg[\left.\frac{\partial^2 L}{\partial M^2}\right|_{\hat{M},\hat{R}}(M-\hat{M})^2 + \left.\frac{\partial^2 L}{\partial R^2}\right|_{\hat{M},\hat{R}}(R-\hat{R})^2 \\ + 2\left.\frac{\partial^2 L}{\partial M \partial R}\right|_{\hat{M},\hat{R}}(M-\hat{M})(R-\hat{R})\Bigg] + \ldots \qquad (4.73) \end{aligned}$$

The quadratic terms enclosed by the brackets in Eqn. 4.73 are those that tell us about the spread of the posterior, and hence the variances on the estimates $\hat{M}, \hat{R}$. These quadratic terms can be written in matrix notation. Let Q be the quadratic term in Eqn. 4.73

$$Q = \begin{pmatrix} M-\hat{M} & R-\hat{R} \end{pmatrix} \begin{pmatrix} A & C \\ C & B \end{pmatrix} \begin{pmatrix} M-\hat{M} \\ R-\hat{R} \end{pmatrix} \qquad (4.74)$$

$$\text{where } A = \left.\frac{\partial^2 L}{\partial M^2}\right|_{\hat{M},\hat{R}} \qquad B = \left.\frac{\partial^2 L}{\partial R^2}\right|_{\hat{M},\hat{R}} \qquad C = \left.\frac{\partial^2 L}{\partial M \partial R}\right|_{\hat{M},\hat{R}} \qquad (4.75)$$

Contours of constant Q are ellipses in M-R space centered at $(\hat{M}, \hat{R})$. The orientation of this error ellipse is determined by the eigenvectors $\mathbf{e_1}$ and $\mathbf{e_2}$ of the matrix

$\left(\begin{smallmatrix} A & C \\ C & B \end{smallmatrix}\right)$. The components (m, r) of eigenvectors $\mathbf{e_1}$ and $\mathbf{e_2}$ satisfy the following equation

$$\begin{pmatrix} A & C \\ C & B \end{pmatrix} \begin{pmatrix} m \\ r \end{pmatrix} = \lambda \begin{pmatrix} m \\ r \end{pmatrix} \tag{4.76}$$

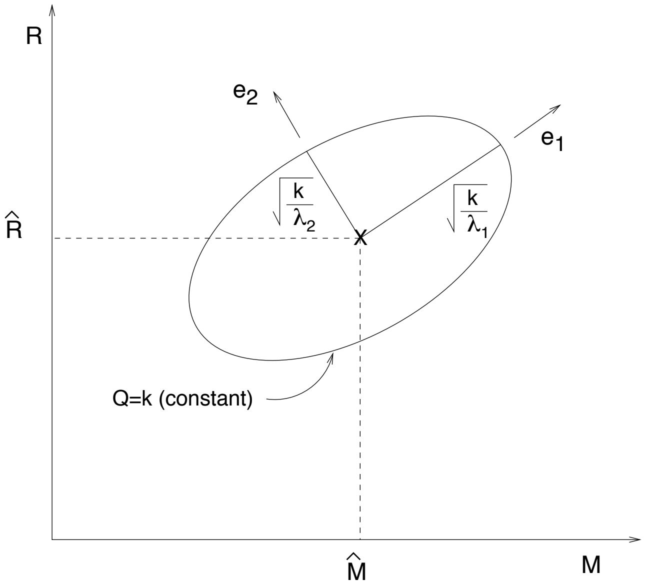


Figure 4.11: Error ellipse for estimates of M, R given available observations, from Sivia (1996).

The eigenvalues, $\lambda_{1,2}$, of the matrix $H = \left(\begin{smallmatrix} A & C \\ C & D \end{smallmatrix}\right)$ of the second derivatives of L, are given by

$$\lambda_{1,2} = \frac{(A+B) \pm \sqrt{(A+B)^2 - 4(AB - C^2)}}{2} \tag{4.77}$$

where A, B, C are as in Eqn. 4.75.

The eigenvectors of $\mathbf{e_{1,2}}$ of H are given by

$$\mathbf{e_{1,2}} = \begin{pmatrix} \frac{A-B\pm\sqrt{(A+B)^2-4(AB-C^2)}}{2C} \\ 1 \end{pmatrix} \tag{4.78}$$

The Taylor series expansion of L about the maxima $(\hat{M}, \hat{R})$ is dominated by the quadratic term Q. Q, which is a function of the second derivatives of L with respect to M, R, gives information regarding the variances on the estimates. The variance on the estimate $\hat{M}$ is

$$\sigma^2_{\hat{M}} = E[(M - \hat{M})^2] = \int\int (M - \hat{M})^2 prob(M, R|Y_A) dMdR \tag{4.79}$$

where 'E[ ]' denotes expectation. In terms of the coefficients of the Taylor series expansion, σ^2_M is

$$\sigma^2_M = \frac{B}{C^2 - AB} \tag{4.80}$$

Similarly, σ^2_R is

$$\sigma^2_R = \frac{A}{C^2 - AB} \tag{4.81}$$

The covariance of M, R, denoted $\sigma^2_{M,R}$ is given by

$$\sigma^2_{M,R} = \int\int (M - \hat{M})(R - \hat{R}) prob(M, R|Y_A) dMdR = \frac{C}{AB - C^2} \tag{4.82}$$

Thus, the second derivative terms of L completely define the variance and covariances of the parameters we are interested in estimating (M, R).

$$\begin{pmatrix} \sigma^2_M & \sigma^2_{M,R} \\ \sigma^2_{M,R} & \sigma^2_R \end{pmatrix} = \frac{1}{AB - C^2} \begin{pmatrix} -B & C \\ C & -A \end{pmatrix} = - \begin{pmatrix} A & C \\ C & B \end{pmatrix}^{-1} = -H^{-1} \tag{4.83}$$

To summarize, Eqn. 4.72 gives the conditions which must be satisfied to obtain the best estimates of $\hat{M}, \hat{R}$ given a single observed amplitude. Eqn. 4.83 gives the covariance matrix for the estimates $(\hat{M}, \hat{R})$. The validity of Eqn. 4.83 depends on the validity of the Taylor series expansion about the maxima identified by Eqn. 4.72.

This holds for the *bivariate* case, where there are two unknowns, M, R. The method above can easily be extended to the *multivariate* case, for example, if the estimation problem is phrased in terms of M, lat_e, lon_e. Let $X_0 = \begin{bmatrix} \hat{M} & lat_e & lon_e \end{bmatrix}$ denote the best magnitude and location estimates, and let $X = \begin{bmatrix} M & lat & lon \end{bmatrix}$ denote the vector of unknowns. The maxima of the posterior density function (or its log_e)

satisfy

$$\left.\frac{\partial L}{\partial X_i}\right|_{X_0} = 0 \tag{4.84}$$

where $i = 1 \cdots 3$. From (Sivia, 1996), the multivariate Taylor series expansion takes the form

$$L = L(X_0) + \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\left.\frac{\partial^2 L}{\partial X_i \partial X_j}\right|_{X_0}(X_i - X_{0i})(X_j - X_{0j}) + \cdots \tag{4.85}$$

As with the bivariate case, the second derivatives of L determine the spread of the posterior, $prob(M, lat, lon|Y)$, and thus the variances of the estimates. The covariance matrix is given by

$$\begin{pmatrix} \sigma_M^2 & \sigma_{M,lat}^2 & \sigma_{M,lon}^2 \\ \sigma_{M,lat}^2 & \sigma_{lat}^2 & \sigma_{lat,lon}^2 \\ \sigma_{M,lon}^2 & \sigma_{lat,lon}^2 & \sigma_{lon}^2 \end{pmatrix} = - \begin{pmatrix} \frac{\partial^2 L}{\partial M^2} & \frac{\partial^2 L}{\partial M \partial lat} & \frac{\partial^2 L}{\partial M \partial lon} \\ \frac{\partial^2 L}{\partial M \partial lat} & \frac{\partial^2 L}{\partial^2 lat} & \frac{\partial^2 L}{\partial lat \partial lon} \\ \frac{\partial^2 L}{\partial M \partial lon} & \frac{\partial^2 L}{\partial lat \partial lon} & \frac{\partial^2 L}{\partial lon^2} \end{pmatrix}^{-1} \tag{4.86}$$

In real-time applications, the derivatives in Eqn. 4.86 can be approximated using finite-difference methods.

Note: Beck and Katafygiotis (1998) have an assymptotic formulation very similar to the discussion of uncertainties in Sivia (1996).

4.6 Summary

In this chapter, I presented a Bayesian approach to the seismic early warning problem. Among the advantages of a Bayesian approach are the use of prior information and the sequential nature of calculations. The Bayes likelihood function is defined in terms of envelope attenuation relationships and linear discriminant analysis magnitude estimators. The types of information that will be used to define the Bayes prior in the following chapters include: Gutenberg-Richter magnitude-frequency relationship, fault locations, previously observed seismicity, and the state of health of the seismic network and the implied geometric constraints. Whether or not to include the

Gutenberg-Richter relationship in the Bayes prior affects the initial source estimates. In the examples that follow, 2 types of source estimates will be tracked: with and without the Gutenberg-Richter in the Bayes prior. How these two estimates might be optimally used by subscribers to early warning alerts is discussed in Chapter 9.

In the sequential analysis of incoming data, the Bayes prior is also the posterior given the available data at the time of the previous estimate. Real-time applications such as seismic early warning can benefit from the sequential nature of Bayesian calculations. At the time of each estimate, it is only necessary to maximize the likelihood of the observations since the previous estimate, and use the posterior from the previous estimate as the prior. This is efficient in terms of necessary calculations, and keeps the minimization problem from growing as a function of time. It is most efficient to formulate the estimation problem in terms of epicentral location (latitude, longitude) rather than epicentral distance. This means there are at most 3 unknowns (not counting the origin time), no matter how many stations or channels are included. It is also more consistent with the form of relevant prior information, such as locations of previously observed seismicity and the Voronoi cells of currently operating stations. Finally, a Taylor series expansion is used to approximate the variances on the source estimates.