Small Problem 8: One-Dimensional Seismic Monitoring

Stuart Russell, Computer Science Division, UC Berkeley
Simplified version. Version 2 11/23/2014.

The goal in this problem is to detect and localize seismic events given signals collected at detector stations over a fixed period.

In this initial version, the world is one-dimensional and the data are simulated. For a complete description of the real problem, see Arora *et al*., (2013). Subsequent versions will include spatial Gaussian process priors for seismicity, velocity, and absorptivity; a 2-D simulated world; and the 3-D real world.

The 1-D world's spatial extent is the unit interval [0,1]. There are five detector stations $s_1, s_2, \ldots, s_5$ with known locations $L(s) = 0, 0.25, 0.5, 0.75, 1.0$.

The generative model has two conceptual parts: the first part is a prior distribution over parameters describing the "physics" of the world, which are fixed but unknown; the second part describes events occurring in a particular episode (a time period of unit length) and any consequent detections. The prior over the physical parameters is as follows:

1. λ_0, the overall rate of event occurrence, is drawn from $\text{Gamma}(\alpha_I, \beta_I)$.
2. Each station has a fixed, unknown background noise level $N(s)$ drawn from an inverse gamma distribution $\text{InvGamma}(\alpha_N, \beta_N)$ and a fixed, unknown false alarm rate $F(s)$ drawn from $\text{Gamma}(\alpha_F, \beta_F)$.
3. The signal velocity is a fixed unknown constant $V_0 = W_0^2$ where W_0 is drawn from $Normal(\mu_V, \sigma_V^2)$. The travel time between two points is $S(x, y) = \frac{|x-y|}{V_0}$.
4. The absorptivity per unit distance is a fixed unknown constant $\alpha_0 = \beta_0^2$ where $\beta_0$is drawn from $\text{Normal}(\mu_B, \sigma_B^2)$.
5. The signal detection capability of each station is governed by fixed, unknown parameters $v(s), \sigma^2(s)$ with $v(s)$ drawn from $\text{Normal}(\mu_v, \sigma_v^2)$ and $\sigma^2(s)$ drawn from an inverse gamma: $\text{InvGamma}(\alpha_s, \beta_s)$.
6. The arrival time measurement error at station s is governed by fixed, unknown parameters $\mu_t(s), \sigma_t^2(s)$ drawn from a normal-inverse-gamma distribution with parameters $(\mu_t, \lambda_t, \alpha_t, \beta_t)$.
7. The amplitude measurement error at station s is governed by fixed, unknown parameters $\mu_a(s), \sigma_a^2(s)$ drawn from a normal-inverse-gamma distribution with parameters $(\mu_a, \lambda_a, \alpha_a, \beta_a)$.

8. The amplitude distribution for noise detections at station s is governed by fixed, unknown parameters $\mu_n(s), \sigma_n^2(s)$ drawn from a normal-inverse-gamma distribution with parameters $(\mu_n, \lambda_n, \alpha_n, \beta_n)$.

Note: We are using the shape/rate parameterization for the Gamma and InvGamma distributions: $Gamma(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$.

For each episode the generative model is as follows:

9. The number of events occurring in the time period [0,1] is drawn from a Poisson with mean λ_0. The location $L(e)$ and the time $T(e)$ of each event are drawn uniformly from [0,1].
10. The magnitude $M(e)$ of each event is drawn from an exponential distribution with rate $\log 10$; i.e., the probability density of magnitude m is proportional to 10^{-m}. The minimum magnitude is 2.0.
11. The arriving log amplitude $A(e,s)$ from event e at station s is given by $M(e) - \alpha_0 |L(e) - L(s)|$.
12. The probability that a signal from event e is detected at a station s is given by $\text{Logistic}\big(A(e,s) - N(s), v(s), \sigma(s)\big)$, where $\text{Logistic}(x, v, \sigma) = 1/\left[1 + \exp\left(-\frac{x-v}{\sigma}\right)\right]$.
13. If event e generates a detection d at station s, the detection attributes are as follows:
    a. The measured arrival time $t(d) = T(e) + S\big(L(e), L(s)\big) + \epsilon_t$ where $\epsilon_t \sim \text{Normal}\big(\mu_t(s), \sigma_t^2(s)\big)$.
    b. The measured log amplitude $a(d) = A(e,s) + \epsilon_a$ where $\epsilon_a \sim \text{Normal}\big(\mu_a(s), \sigma_a^2(s)\big)$.
    c. The sign $s(d)$ indicating the direction of arrival of the signal is $+1$ if $L(e) < L(s)$, and -1 otherwise.
14. The number of noise detections d at station s is $\text{Poisson}\big(N(s)\big)$:
    a. The measured arrival time $t(d)$ is uniform in [0,1].
    b. The measured log amplitude $a(d) \sim \text{Normal}\big(\mu_n(s), \sigma_n^2(s)\big)$.
    c. The sign $s(d)$ indicating the direction of arrival of the signal is drawn uniformly from $\{+1, -1\}$.
    d. Note:

The file "`problem-8-episodes100.data`" contains the scenarios. We will use the first 75 episodes for training and the last 25 episodes for testing. We will treat each testing episode as a query, so there are 25 queries. For each episode, we provide the detections, the true events, and the associations between them. When an episode is being used for testing, the true events and all of the associations are the desired outputs, and should obviously not be used for training. The desired output for each period is a *bulletin*, i.e., a set of events with times, locations, and magnitudes.

To evaluate a bulletin, a matching is constructed between the bulletin and the ground truth for that period (the code for this is provided as `eval1d.py` and `mwmatching.py` in the `problem-8-files` folder) . A bipartite graph is created between predicted and true events. An edge is added between a predicted and a true event that is at most d_{max} in distance and t_{max} in time apart. We will use $d_{max} = t_{max} = 0.01$, although these may need to be adjusted based on experience. The weight of the edge is the distance between the two events. Finally, a minimum weight–maximum cardinality matching is computed on the graph. Three metrics are extracted from this matching:

- Metric 1: Precision (percentage of predicted events that are matched);
- Metric 2: Recall (percentage of true events that are matched), and
- Metric 3: Average error (average distance between matched events).

The fixed, known parameters of the model are marked in red above. The values used for the current data set are as follows:

$\alpha_I = 20$ $\beta_I = 2$
$\alpha_N = 3$ $\beta_N = 2$
$\alpha_F = 20$ $\beta_F = 1$
$\mu_V = 5$ $\sigma_V^2 = 1$
$\mu_B = 2$ $\sigma_B^2 = 1$
$\mu_v = 1$ $\sigma_v^2 = 1$
$\alpha_s = 2$ $\beta_s = 1$
$\mu_t = 0$ $\lambda_t = 1000$ $\alpha_t = 20$ $\beta_t = 1$
$\mu_a = 0$ $\lambda_a = 1$ $\alpha_a = 2$ $\beta_a = 1$
$\mu_n = 0$ $\lambda_n = 1$ $\alpha_n = 2$ $\beta_n = 1$