

Summary: Bayes Linear for Avalanches

The basic system can be written

$$\mathbf{y} = \mathbf{f}(\mathbf{x}, \theta) + \epsilon \quad (1)$$

where \mathbf{y} are the dependent variables, \mathbf{f} encodes the physics in a system of PDEs or nonlinear equations, \mathbf{x} are input parameters that are fixed (either because of knowledge, or out of total ignorance), θ are inputs about which we make statements of prior knowledge or belief, and ϵ represents modeling error.

For the granular avalanche problem,

$$\mathbf{y} = \begin{pmatrix} h \\ hu \\ hv \end{pmatrix} \quad (2)$$

Here h is the thickness of the flowing layer of material, and (u, v) are the velocities. The physics is represented by PDEs for the conservation of mass and momentum (i.e., \mathbf{f} are PDEs in 2 space dimensions plus time, with input data \mathbf{x}, θ). The inputs \mathbf{x} include DEM terrain data (at a specified ‘resolution’). The stochastic inputs $\theta = (\phi_{\mathbf{bed}}, \phi_{\mathbf{int}}, \mathbf{M}, \theta_{\mathbf{r}})^{\mathbf{T}}$ where ϕ_{bed} is the bed friction angle - which may be a function of position, or regions of different bed friction, ϕ_{int} is the internal friction angle which is constant in a given run, but for which a range of angles should be examined, M is the initial volume of material, and θ_r includes all the other variable inputs such as the location of the initial mass (2 coordinate locations), or its initial velocity (2 components).

We wish to develop 3 features of work:

1. Experimental Design - how to choose the θ s to best fill out the design space and see the important features of the response surface
2. Calibration - how to use field and lab observations to inform what we know about the model
3. Forecast - how to use the model to make predictions of flow path and depth and areas at risk, and to place quantitative bounds on these predictions.

Comments:

1. For experimental design, we can run numerical experiments. Very accurate runs (fine grid) take several hours up to days on 32 or more processors; less accurate runs (coarse grids) take 1-4 hrs on one or a few processors. We also have available fast runs using computational models based on other equations, or more approximate physics.
2. Regarding calibration, geologists can estimate the spread and depth of deposits left by past events. These observations are made by drilling cores to measure depth at several locations in the deposit region, and are subject to error themselves. In some cases, geo-scientists can estimate the velocity of a past flow at a few select locations along the flow path, for example by seeing how far up a valley wall the flow runs up as it curves. Core samples may indicate whether the flow bulked through erosion.
3. Finally, forecasting must make some statement about the limitations of the model, physics that is not included, etc.

To proceed, run the simulation at reasonable grid refinement for several inputs θ chosen, for example, by a Latin hypercube sampling. This defines the prior probability distribution. We would like to choose the next input point based on a Bayes linear approximation. So this means forming an approximate (polynomial) response surface

$$\mathbf{f}^{\mathbf{p}} = \sum \alpha \theta^{\mathbf{p}} + e(\theta, \alpha)$$

where α is a multi-indexed coefficient, the $\theta^{\mathbf{p}}$ is a multi-index of the inputs θ raised to powers designated by the multi-index \mathbf{p} . [For example, the GMFG used 2 inputs θ_1, θ_2 , and adaptively built up the polynomial approximation to 5th order, choosing the next input point by minimizing the expected variance (i.e. choosing the point with the largest variance currently). There is some clustering of sample points especially as one changes the order of the approximating polynomial. A question is how to choose e – is there a basis for, say, choosing it as a multi-variate normal with 0 mean and some variance? Why choose one metric over another? For example, why minimize the expected variance $E(var)$ rather than maximum error $\max(var)$? How to choose ‘good’ models for e ? [Note: Dalbey performed some 10,000 runs a reasonably fine grid, largely covering the input space (LHS). This serves as a ‘truth’ surface, and we can measure how far off any prediction is from this surface.]

Next comes calibration. We have a set of observations Z , which typically include depth (i.e., h), maybe some velocity estimates at certain locations as mentioned above, and a spread of the final deposit (that is $(x, y) : h(x, y, t_{final}) > tol$). Observation errors can be large, and there is little but guessing to inform the nature of these errors. We have an approximate response surface, so given an observation we could use \mathbf{f} to write the probability $P(z|\theta)$. Turn this around to calibrate the approximate response surface using Bayes Theorem

$$P(\theta|z) = P(z|\theta)P(\theta)/P(z)$$

How does one assign the probabilities to θ and z ? We have some information - for example, there is evidence of the distribution of the frequency of a flow of given size, so this is a good $P(\theta)$ for that one input component. We might use a distribution of friction angles for different materials as the prior for the friction angle. But how likely is a flow? What is $P(z)$?

Now to forecasting. Given an improved p.d.f. for θ , propagate this distribution through \mathbf{f} to obtain an improved p.d.f. for \mathbf{y} . This could be done by Polynomial Chaos or, better, PCQ. That gives a hazard map - in principle, we could assign the probability of a specific location ever seeing a flow of a given depth (conditioned on a flow happening!). Importantly, how to model ϵ ? For example, the vanilla model does not include a fluid phase, nor erosion. Does it help for hazard assessment to join TITAN modeling with, for example, LaharZ, which makes its own set of modeling errors, the errors of each approach complementing each other? How might we include the errors in the topography (here there are known unknowns, i.e. specified errors in the measurements such as the rms error of measurements, and unknown unknowns such as the x , y -errors evidenced in Sheridan's work at Banos)? How to model the error in $\mathbf{f}^{expensive} - \mathbf{f}^{cheap}$?