Stat243: Problem Set 5, Due Dec. 4

November 18, 2013

This covers material in Units 12 and 13 on simulation and optimization.

It's due at the start of class on Wed., Dec. 4.

Please follow the guidelines we have discussed previously on formatting your answers. Solutions that are not formatted will have points taken off. **Also note that handwritten solutions to the non-computer problems are fine, so long as your handwritten solutions are easy to follow. If you do have some handwritten solutions, in your paper submission, please have your answers to problems be in order, by stapling together printed and handwritten pages as needed.**

Please note my comments in the syllabus about when to ask for help and about working together.

Problems

1. The simulation study discussion questions you answered for section Nov. 4. You do not need to do anything further.

2. Let's consider importance sampling and explore the need to have the sampling density have heavier tails than the density of interest. Assume that we want to estimate EX and $E(X^2)$ with respect to a density, f. We'll make use of the Pareto distribution, which has the pdf $p(x) = \frac{\beta\alpha^\beta}{x^{\beta+1}}$ for $\alpha < x < \infty$, $\alpha > 0$, $\beta > 0$. The mean is $\frac{\beta\alpha}{\beta-1}$ for $\beta > 1$ and non-existent for $\beta \leq 1$ and the variance is $\frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}$ for $\beta > 2$ and non-existent otherwise.

   (a) Does the tail of the Pareto decay more quickly or more slowly than that of an exponential distribution?

   (b) Suppose f is an exponential density with parameter value equal to 1, shifted by two to the right so that $f(x) = 0$ for $x < 2$ and our sampling density, g, is a Pareto distribution with $\alpha = 2$ and $\beta = 3$. Use $m = 10000$ to estimate EX and $E(X^2)$. Recall that $\text{Var}(\hat{\mu}) \propto \text{Var}(h(X)f(X)/g(X))$. Create histograms of $h(x)f(x)/g(x)$ and of the weights $f(x)/g(x)$ to get an idea for whether $\text{Var}(\hat{\mu})$ is large. Note if there are any extreme weights that would have a very strong influence on $\hat{\mu}$.

   (c) Now suppose f is the Pareto distribution described above and our sampling density, g, is the exponential described above. Respond to the same questions as for part (b).

3. Consider probit regression, which is an alternative to logistic regression for binary outcomes. The probit model is $P(Y_i = 1) = \Phi(X_i^\top \beta)$ where Φ is the standard normal CDF. We can rewrite this model with latent variables, one latent variable for each observation:

$$\begin{aligned} y_i &= I(z_i > 0) \\ z_i &\sim \mathcal{N}(X_i^\top \beta, 1) \end{aligned}$$

(a) Design an EM algorithm to estimate β, taking the complete data to be $\{Y, Z\}$. You'll need to make use of $E(W|W > \tau)$ and Var$(W|W > \tau)$ where W is normally distributed. Be careful that you carefully distinguish β from the current value at iteration t, β^t, in writing out the expected log-likelihood and computing the expectation and that your maximization be with respect to β (not β^t). Also be careful that your calculations respect the fact that for each z_i you know that it is either bigger or smaller than 0 based on its y_i. You should be able to analytically maximize the expected log likelihood. A couple hints:

i. From the Johnson and Kotz bibles on distributions, the mean and variance of the truncated normal distribution, $f(w) \propto \mathcal{N}(w; \mu, \sigma^2) I(w > \tau)$, are:

$$\begin{aligned} E(W|W > \tau) &= \mu + \sigma\rho(\tau^*) \\ V(W|W > \tau) &= \sigma^2 \left(1 + \tau^*\rho(\tau^*) - \rho(\tau^*)^2\right) \\ \rho(\tau^*) &= \frac{\phi(\tau^*)}{1 - \Phi(\tau^*)} \\ \tau^* &= (\tau - \mu)/\sigma, \end{aligned}$$

where $\phi(\cdot)$ is the standard normal density and $\Phi(\cdot)$ is the standard normal CDF.

ii. You should recognize that your expected log-likelihood can be expressed as a regression of some new quantities (which you might denote as m_i, $i = 1, \ldots, n$, that are functions of β^t and y_i) on X.

(b) Propose reasonable starting values for β.

(c) Write an R function, with auxiliary functions as needed, to estimate the parameters. Make use of the initialization from part (b). You may use *lm()* for the update steps. You'll need to include criteria for deciding when to stop the optimization. Test your function using data simulated from the model, with say $\beta_0, \beta_1, \beta_2, \beta_3$. Take $n = 100$ and the parameters such that with complete data, $\hat{\beta}_1/se(\hat{\beta}_1) \approx 2$ and $\beta_2 = \beta_3 = 0$.

(d) A different approach to this problem just directly maximizes the log-likelihood of the observed data. Estimate the parameters (and standard errors) for your test cases using *optim()* with the BFGS option in R. Compare how many iterations EM and BFGS take.

4. The following is a real optimization problem in a project I've been working on, collaborating with an astrophysicist and computer scientist at LBL (the national lab just up the hill from campus) as well as a colleague here in the Statistics Department. The data are the public spectrophotometric time series of the Type Ia supernova SN 2011fe from multiple nights worth of measurements. The basic context is measuring flux from the supernova as a function of (log) wavelength and time [also called 'day' or 'phase' in the dataset]. The goal in the project is basically to estimate the flux as a smooth function (i.e. 2-d surface) of (log) wavelength and time based on the pairs{wavelength, time} for which we have data. To do this, there are two steps:

- Estimate the parameters of a statistical model for the data, which in this case is a Gaussian process model that gives us a multivariate normal distribution for the data.
- Use the estimated parameters to interpolate to {wavelength,time} values for which we do not have data (I'm leaving out the details of how this is done.). By doing so on a grid of {wavelength,time} values, we can plot an estimate of the flux surface as a function of wavelength and time.

Your task here is to carry out (a) based on the normal distribution below given below. To make the problem computationally tractable without parallel processing, you'll use data from only a subset of wavelengths and time points to allow for optimization on a single machine (we're using parallel processing in the real project). We assume a somewhat complicated normal distribution for the observations that accounts for correlation between the fluxes at nearby wavelengths and times. The model explains some of the variability in the observations through the mean of the normal distribution and some through the covariance structure.

The normal likelihood, as a function of parameters, $\theta = \{\kappa, \lambda, \sigma^2, \rho_w, \rho_t, \tau^2, \alpha\}$, is

$$Y \sim \mathcal{N}(\mu_\theta, C_\theta)$$

with likelihood function proportional to

$$|C|^{-1/2} \exp\left(-\frac{1}{2}(Y-\mu_\theta)^\top C_\theta^{-1}(Y-\mu_\theta)\right)$$

where μ_θ is a vector (with one value per observation) such that the value for observation Y_{wt} at log wavelength w and time t is

$$E(Y_{wt}) = \mu(t;\kappa,\lambda) = \kappa f(\frac{t}{\lambda})$$

where $f(\cdot)$ is a fixed smooth (spline) function of time t that represents the average effect (across many different supernova) of time on flux. κ and λ allow this "template" function, f, to be stretched and scaled to better represent the effect of time for a particular supernova. The function $\mu(t;\kappa,\lambda)$ is coded for you and available in *ps5prob4.R*. C_θ is a matrix containing all the pairwise covariances between the observations, such that for an observation at log wavelength w and time t and an observation at log wavelength w' and time t' the entry in the covariance matrix is

$$\text{Cov}(Y_{wt}, Y_{w't'}) = \sigma^2 \exp\left(\frac{-|w-w'|}{\rho_w}\right) \cdot \exp\left(\frac{-|t-t'|}{\rho_t}\right) + \tau^2 I(t=t') + \alpha v_{wt}^2 I(w=w', t=t')$$

In this expression, the two exponential functions stipulate that the covariance decays exponentially as a function of the lags in the log wavelength and time domains. The $\tau^2 I(t=t')$ adds a variance component for observations that occur at the same time (day), reflecting the fact that all the observations (over multiple wavelengths) at a given time may be systematically shifted relative to values at other times, while the $\alpha v_{wt}^2 I(w=w', t=t')$ term adds a measurement error component along the diagonal of C_θ, with a fixed observed error, v_{wt}^2, for each observation, scaled by α. $I(\cdot)$ is the indicator function that evaluates to 1 when its argument is true and 0 otherwise. The data are contained in the *data* object in the file *ps5prob4.RData*, with *fluxerror* being the v_{wt} values. Don't forget to use the log of wavelength and the square of the v_{wt} values.

Address the following questions in your solution:

(a) Plot the data with respect to log wavelength and time to get a sense for how flux varies as a function of the variables.

(b) Decide on a reasonable set of initial values based on the scale of the flux values and the log wavelength and time values. Here are some suggestions:

- For κ and λ, plot $\mu(t;\kappa,\lambda)$ for some values of κ and λ to see what pair of values might give a mean function that more or less matches how flux varies with time, averaging over different wavelengths.

- The overall variability of the data should roughly match $\text{Var}(Y_{wt}) = \text{Cov}(Y_{wt}, Y_{wt})$. That is, $\sigma^2 + \tau^2 + \alpha\bar{v}_{\cdot\cdot}^2$ should roughly match the overall variability of the data. τ^2 should roughly match the variability of $\bar{Y}_{\cdot t}$ across the different time periods.
- When ρ_w gets close to zero (e.g., when the values are smaller than the gaps in log wavelength between "nearby" observations), then this says that there is little relationship between the flux values for two observations with similar log wavelengths. When ρ_w gets large (e.g., larger than the difference between the minimum and maximum log wavelengths in the dataset), then this says that the flux at the minimum and flux at the maximum log wavelengths are increasingly similar. Analogous reasoning holds for ρ_t.

(c) Write a log-likelihood function in R that you can use for the optimization, keeping in mind our best practices from the linear algebra unit.

(d) Use R's optimization tools to optimize the log likelihood and find the maximum likelihood estimate and asymptotic standard errors for θ. Optimization will likely take at least a few minutes; you may want to increase the convergence tolerance in doing initial exploratory optimizations to reduce the time involved. In doing the optimization, address the following:

- Reparameterize to avoid constrained parameters.
- Consider a variety of starting values to boost your confidence that you have found the global maximum.
- Consider two or three optimization methods to boost your confidence that you have found the global maximum. Note whether some are quicker than others and whether some do not reach the best of the optima you find.
- When using *optim()*, consider whether using *parscale* in your optimization might be worthwhile.

Be wary of any of the estimates getting very large or very small, as this may indicate convergence to a local optimum in an extreme part of the parameter space. However, you may find that the MLE has some of the variance components (i.e., σ^2, τ^2, α) near zero.

I realize that this statistical model may not seem intuitive and am happy to describe the problem further if you ask me, to help you better understand the context for the optimization.