

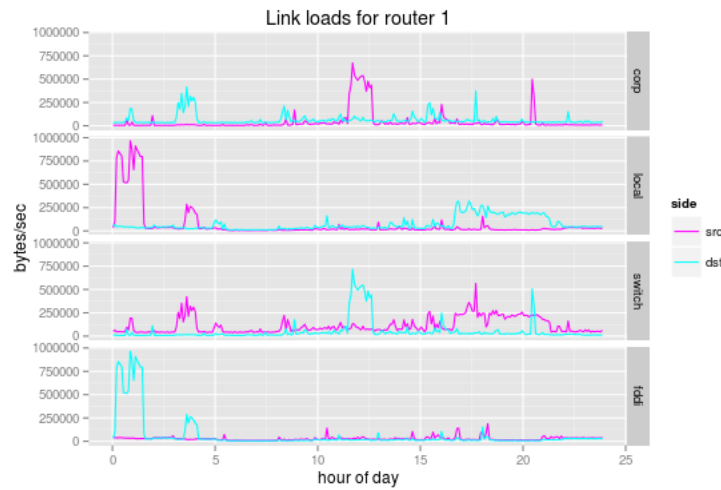
# STAT-221: Pset 5

KEVIN KUATE FODOUOP  
Harvard University

## Abstract

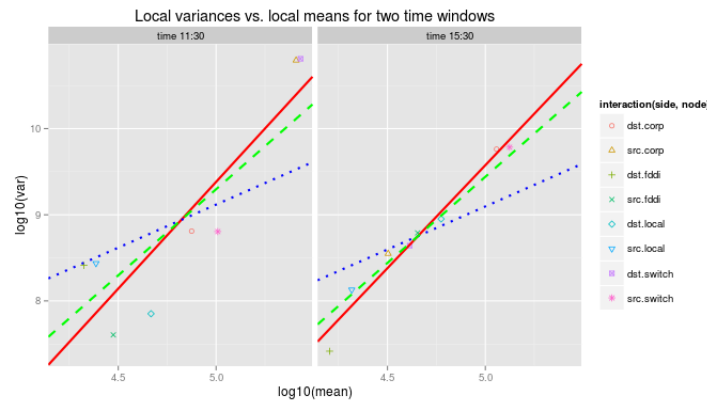
In this homework we use an implementation of the Expectation-Maximization (EM) algorithm to lead inference on non-observable origin-destination (OD) flows in a communication network where only link loads are measured. Measurements are taken every five minutes. Two implementation of EM are derived, replicating the two models described in Cao et al. (JASA, 2000).

question 1.1 We replicate figure 2 of the paper in figure 1, using link loads from `1router_allcount.dat`.



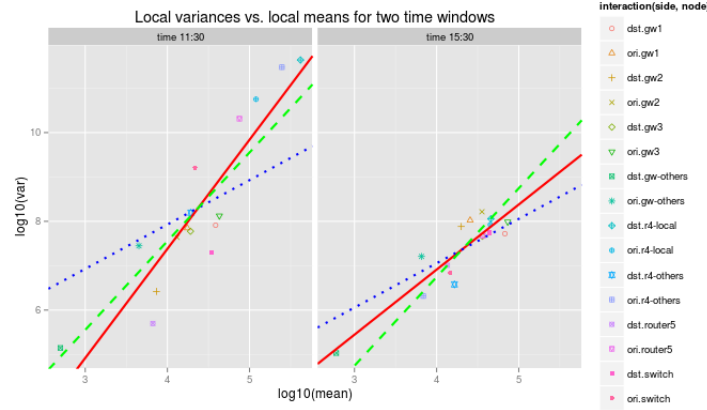
**Figure 1:** Link loads on different node of router 1's subnetwork, replicating Cao et al. figure 2.

question 1.2 We replicate figure 4 of the paper with `1router_allcount.dat`, of log variance against log mean in two time windows of 55 minutes. Linear fits are plotted for unfixed slope and slopes fixed at  $c = 1$  and  $c = 2$ . From the plot  $c = 2$  seems to give better regression results.



**Figure 2:** Local variances versus local means on log scale for the first data set. Linear regression in red,  $c = 1$  in dashed blue and  $c = 2$  in dashed green.

Same figure is plotted for `2router_linkcount.dat` on figure 3. In this dataset there are 8 different type of nodes, so 16 combinations of side (origin or destination) - node. Again  $c = 2$  seems to give better results than  $c = 1$ .



**Figure 3:** Local variances versus local means on log scale for the second data set. Linear regression in red,  $c = 1$  in dashed blue and  $c = 2$  in dashed green.

### question 1.3

We model the  $I$  unobserved OD counts  $x_t$  at time  $t$  as a vector of independent normal random variables

$$x_t \sim \text{Normal}(\lambda, \Sigma)$$

with  $\Sigma = \phi \text{diag}(\sigma^2(\lambda_1), \dots, \sigma^2(\lambda_I))$ , where  $\sigma^2(\lambda) = \lambda^c$ . And the observed link byte counts  $y_t$  as

$$y_t = Ax_t \sim \text{Normal}(A\lambda, A\Sigma A')$$

We base inference on maximum likelihood on iid measurement of this distribution. There is no closed form solution to the likelihood maximization, so that an EM algorithm is implemented to find the parameter solution.

We do not assume a particular value for  $c$ , and our parameter is  $\theta = (\lambda, \phi)$  ( $16 + 1 = 17$  dimensional).

The EM conditional expectation function  $Q$  is

$$Q(\theta, \theta^{(k)}) = E_q [\log(p(y, x|\theta))]$$

With  $q = p(x|y, \theta^{(k)})$ , so that

$$\begin{aligned} Q(\theta, \theta^{(k)}) &= E \left[ \log(p(Y, X|\theta)) | Y, \theta^{(k)} \right] \\ &= E \left[ \log(p(X|\theta)) | Y, \theta^{(k)} \right] \\ &= E \left[ l(\theta|X) | Y, \theta^{(k)} \right] \end{aligned}$$

With  $l(\theta|X)$  latent variable likelihood. We have

$$\begin{aligned} l(\theta|X) &= -\frac{T}{2} \log|\Sigma| - \frac{1}{2} \sum_{t=1}^T (x_t - \lambda)' \Sigma^{-1} (x_t - \lambda) \\ &= -\frac{T}{2} \log|\Sigma| - \frac{1}{2} \sum_{t=1}^T x_t' \Sigma^{-1} x_t - \frac{1}{2} \sum_{t=1}^T x_t' \Sigma^{-1} \lambda - \frac{1}{2} \sum_{t=1}^T \lambda' \Sigma^{-1} x_t - \frac{1}{2} \sum_{t=1}^T \lambda' \Sigma^{-1} \lambda \end{aligned}$$

So that, taking expectation given  $(Y, \theta^{(k)})$

$$\begin{aligned} Q(\theta, \theta^{(k)}) &= -\frac{T}{2} \log|\Sigma| - \frac{1}{2} \sum_{t=1}^T \left( \text{tr}(\Sigma^{-1} \text{var}(x_t|Y, \theta^{(k)})) + E(x_t'|Y, \theta^{(k)}) \Sigma^{-1} E(x_t|Y, \theta^{(k)}) \right) \\ &\quad - \frac{1}{2} \sum_{t=1}^T E(x_t'|Y, \theta^{(k)}) \Sigma^{-1} \lambda - \frac{1}{2} \sum_{t=1}^T \lambda' \Sigma^{-1} E(x_t|Y, \theta^{(k)}) - \frac{1}{2} \sum_{t=1}^T \lambda' \Sigma^{-1} \lambda \end{aligned}$$

Using the variance formula for a quadratic form to compute the second term,  $E(\epsilon' \Lambda \epsilon) = \text{tr}(\Lambda \epsilon) + \mu' \Lambda \mu$ . As  $x_t$  is only dependent on  $y_t$ , the conditional expectation function simplifies to

$$Q(\theta, \theta^{(k)}) = -\frac{T}{2} \left( \log|\Sigma| + \text{tr}(\Sigma^{-1} R^{(k)}) \right) - \frac{1}{2} \sum_{t=1}^T (m_t^{(k)} - \lambda)' \Sigma^{-1} (m_t^{(k)} - \lambda)$$

Where we have conditional mean and variance of  $x_t$

$$\begin{aligned} m_t^{(k)} &= E(x_t|y_t, \theta^{(k)}) \\ &= \lambda^{(k)} + \Sigma^{(k)} A' (A \Sigma^{(k)} A')^{-1} (y_t - A \lambda^{(k)}) \\ R^{(k)} &= \text{var}(x_t|y_t, \theta^{(k)}) \\ &= \Sigma^{(k)} - \Sigma^{(k)} A' (A \Sigma^{(k)} A')^{-1} A \Sigma^{(k)} \end{aligned}$$

Those expression are derived by considering the multivariate normal vector  $(x_t, y_t)$ , and apply formulas for projected multivariate normals given that  $\text{cov}(x_t, y_t) = \text{cov}(x_t, A x_t) = A \Sigma^{(k)}$  and  $\text{cov}(y_t, y_t) = A \Sigma^{(k)} A'$  given  $\theta^{(k)}$ .

Expanding the expression of  $Q$ , we have

$$\begin{aligned} Q(\theta, \theta^{(k)}) &= -\frac{T}{2} \sum_{i=1}^I c \log(\lambda_i) - \frac{T}{2} \sum_{i=1}^I \frac{r_{ii}^{(k)}}{\phi \lambda_i^c} - \frac{1}{2} \sum_{t=1}^T \sum_{i=1}^I \frac{(m_{t,i}^{(k)})^2}{\phi \lambda_i^c} \\ &\quad + \sum_{t=1}^T \sum_{i=1}^I \frac{\lambda_i m_{t,i}^{(k)}}{\phi \lambda_i^c} - \frac{1}{2} \sum_{t=1}^T \sum_{i=1}^I \frac{\lambda_i^2}{\phi \lambda_i^c} \end{aligned}$$

So that taking  $\frac{\partial Q}{\partial \theta}$  gives us the system of equations (first I ones due to  $\lambda$ , last one to  $\phi$ )

$$c \phi \lambda_i^c + (2 - c) \lambda_i^2 - 2(1 - c) \lambda_i b_i^{(k)} - c a_i^{(k)} = 0 \quad i = 1 \dots I \quad (1)$$

$$\sum_{i=1}^I \lambda_i^{-c+1} (\lambda_i - b_i^{(k)}) = 0 \quad (2)$$

Where we have defined

$$a_i^{(k)} = r_{ii}^{(k)} + \frac{1}{T} \sum_{t=1}^T (m_{t,i}^{(k)})^2$$

$$b_i^{(k)} = \frac{1}{T} \sum_{t=1}^T m_{t,i}^{(k)}$$

We hence derive the steps of the EM algorithm for our iid model.

**Data:** Observed link loads  $Y$ .

**Result:** MLE of parameter  $\theta$ .

initialization:  $\theta = \theta_0$  positive parameter.

**while**  $|Q(\theta^{(k+1)}, \theta^{(k)}) - Q(\theta^{(k)}, \theta^{(k-1)})| > \epsilon$  **do**

    - Update step:  $k = k + 1$

    - E-step: Compute  $m_t^{(k)}, R^{(k)}, a_i^{(k)}, b_i^{(k)}$ .

    - M-step **if**  $c = 1$  or  $c = 2$  **then**

        1. Solve (1) for  $\lambda$  analytically given  $\phi$  (positive solution).

        2. Solve for  $\phi$  using fractional steps Newton-Raphson or constrained optim optimization (ensures  $\phi$  positive).

**else**

        Update  $\theta$  using constrained optim optimization (ensures parameters positive).

**end**

**end**

**Algorithm 1:** EM algorithm for iid model

*question 1.4* We fix the variance and window frame paramters to be  $c = 2$  (most appropriate according to exploratory data analysis),  $w = 11$ . The MLE equations (1) then becomes

$$\phi \lambda_i^2 + \lambda_i b_i^{(k)} - a_i^{(k)} = 0 \quad i = 1 \dots I$$

Which gives us a positive solution (the biggest of the two roots of the equation) for  $\lambda_i$  given  $\phi$

$$\lambda_i^{(k+1)} = \frac{\sqrt{(b_i^{(k)})^2 + 4\phi^{(k)} a_i^{(k)}} - b_i^{(k)}}{2\phi^{(k)}}$$

Setting  $\lambda$  to those values, we have  $f_i(\theta) = 0$  for  $i = 1 \dots I$  where  $f(\theta)$  is the left hand-side of equations (1) and (2), so that the one-step Newton-Raphson algorithm

$$\theta^{(k+1)} = \theta^{(k)} - \left[ F(\theta^{(k)}) \right]^{-1} f(\theta^{(k)})$$

Reduces to

$$\phi^{(k+1)} = \phi^{(k)} - \left( \left[ F(\theta^{(k)}) \right]^{-1} \right)_{I+1, I+1} f_{I+1}(\theta^{(k)})$$

With

$$f_{I+1}(\theta^{(k)}) = \sum_{i=1}^I \frac{\lambda_i - b_i^{(k)}}{\lambda_i}$$

And the Jacobian  $F(\theta^{(k)})$  defined by

$$\begin{aligned}\frac{\partial f_i}{\partial \lambda_j} &= \delta_{ij} \left( \frac{4\phi}{\lambda_i} + 2b_i^{(k)} \right) \\ \frac{\partial f_{I+1}}{\partial \lambda_j} &= \frac{b_j^{(k)}}{\lambda_j} \\ \frac{\partial f_i}{\partial \phi} &= 2\lambda_i^2 \\ \frac{\partial f_{I+1}}{\partial \phi} &= 0\end{aligned}$$

To solve the M-step of our EM algorithm, we adopt an iterative fractional Newton-Raphson method, by dividing the step in the former update of  $\phi$  by bigger and bigger integer if the update is negative. However this approach failed to converge fast enough (steps are too small), and we falled back on an optim optimization method. We extend the basic iid model to a local iid model,

by setting time moving windows in which observations are treated iid.

$$y_{t-h}, \dots, y_{t+h} \sim \text{Normal}(A\lambda_t, A\Sigma_t A')$$

With  $\Sigma_t = \phi_t \text{diag}(\lambda_t^2)$ . On each of those windowed data set, the MLE parameter  $\theta_t$  is fitted using the previously derived iid model EM.

*question 1.5* We derive the EM algorithm for the refined model of the paper's fourth section. This refined model adds smoothing to the estimates by considering the parameters as following hidden markov model, resulting in an algorithm similar to the Kalman Filter.

$\eta_t = (\log(\lambda_t), \log(\phi_t))$  is modeled as a random walk state

$$\eta_t = \eta_{t-1} + v_t$$

With  $v_t \sim \text{Normal}(0, V)$ , and  $V$  fixed variance matrix. We denote all the current information until time  $t + h$  by  $\tilde{Y}_t = (y_1, \dots, y_{t+h})$ , and the window data set at time  $t$   $Y_t$ . The goal of our EM algorithm is to find the MAP estimator, ie. the mode of the posterior distribution  $p(\eta_t | \tilde{Y}_t)$ .

We actually have

$$p(\eta_t | \tilde{Y}_t) = p(\eta_t | \tilde{Y}_{t-1}, Y_t) \propto p(\eta_t | \tilde{Y}_{t-1}) p(Y_t | \eta_t)$$

Using Baye's rule. Hence we now want to maximize the log posterior distribution (removing additive constants)

$$g(\eta_t) = \log(\pi_t(\eta_t)) + \log(p(Y_t | \eta_t))$$

Where  $\pi_t(\eta_t) = p(\eta_t | \tilde{Y}_{t-1})$  prior at time  $t$ , and  $\log(p(Y_t | \eta_t)) = \log(p(Y_t | \theta_t))$  is what we were trying to maximize in the previous local EM algorithm.

This modifies our EM algorithm by adding a constant (in that it does not depend on the data) prior component to the  $Q$  expectation function, ie.

$$Q(\theta, \theta_t^{(k)}) = Q_{local}(\theta, \theta_t^{(k)}) + \log(\pi_t(\log(\theta)))$$

Hence we just need to come up with the expression of  $\log(\pi_t(\eta_t))$  to modify our E-step, and optimize the new  $Q$  for our M-step.

Integrating on all values of former step states, we have

$$\pi_t(\eta_t) = \int p(\eta_{t-1} | \tilde{Y}_{t-1}) p(\eta_t | \eta_{t-1}) d\eta_{t-1}$$

With  $\eta_t | \eta_{t-1} \sim \text{Normal}(\eta_{t-1}, V)$ . The paper approximates the former state posterior as a normal distribution, ie.  $\eta_{t-1} | \tilde{Y}_{t-1} \sim \text{Normal}(\hat{\eta}_{t-1}, \hat{\Sigma}_{t-1})$ , where  $\hat{\eta}_{t-1}$  is the previous state posterior mode and  $\hat{\Sigma}_{t-1} = \ddot{g}(\hat{\eta}_t)^{-1}$  (local normal approximation). So the integral for the prior gives us

$$\pi_t(\eta_t) = \text{Normal}(\hat{\eta}_{t-1}, \hat{\Sigma}_{t-1} + V)$$

So that we obtain a  $Q$  expectation function

$$Q(\theta, \theta_t^{(k)}) = Q_{local}(\theta, \theta_t^{(k)}) + \log(\text{dnorm}(\log(\theta), \hat{\eta}_{t-1}, \hat{\Sigma}_{t-1} + V)) \quad (3)$$

And the refined model EM algorithm proceeds as follows:

**Data:** Observed link loads  $Y$ .

**Result:** MAP of parameters  $\theta_t$ .

**for**  $t=1$  **to**  $T$  **do**

    initialization:  $\theta_t = \theta_0$  positive parameter.

**while**  $|Q(\theta_t^{(k+1)}, \theta_t^{(k)}) - Q(\theta_t^{(k)}, \theta_t^{(k-1)})| > \epsilon$  **do**

        - Update step:  $k = k + 1$

        - E-step: Compute  $m_t^{(k)}, R^{(k)}, a_i^{(k)}, b_i^{(k)}, \hat{\eta}_{t-1}, \hat{\Sigma}_{t-1}$  (using numerical methods for the Hessian).

        - M-step: Update  $\theta_t$  using constrained optim optimization (ensures parameters positive) on the  $Q$  function defined in (3).

**end**

**end**

**Algorithm 2:** EM algorithm for refined model