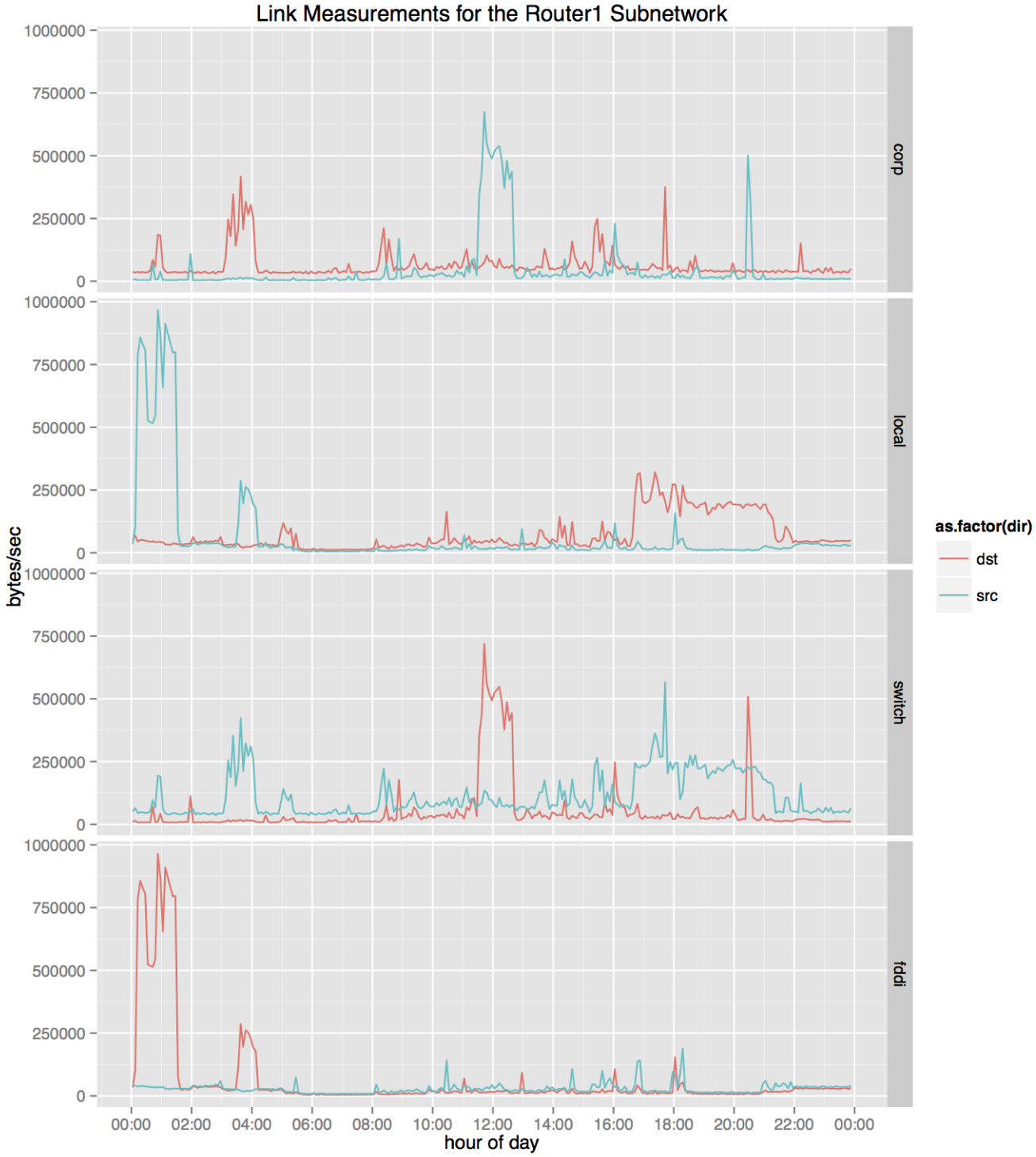
Stat 221 Problem Set 5

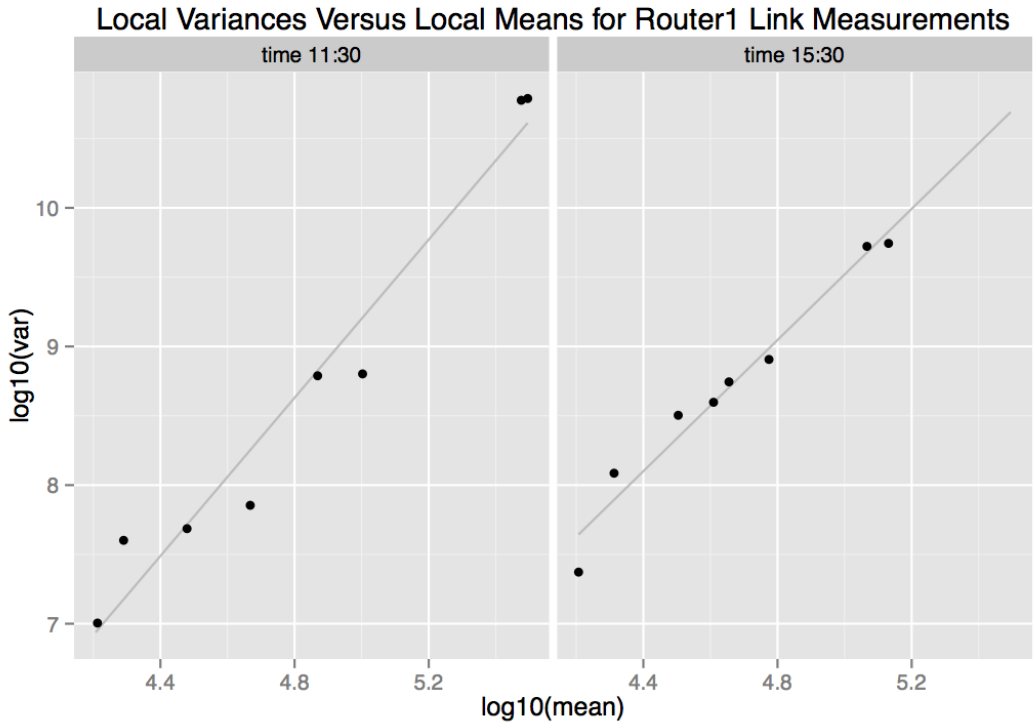
Albert Young and Marco Gentili

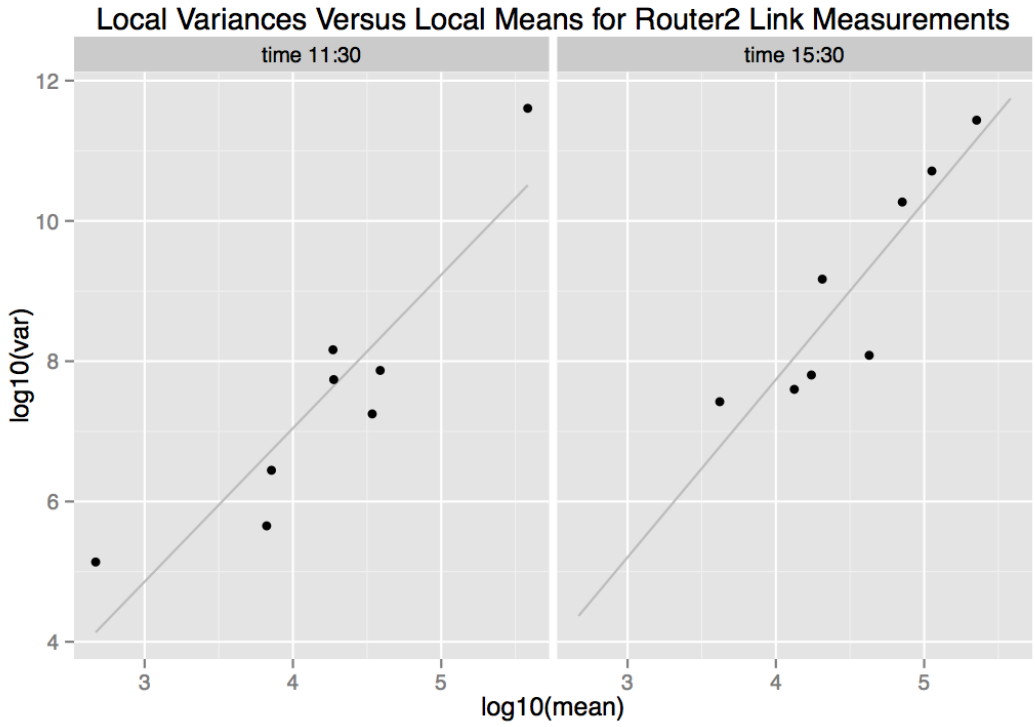
18 November 2014

Question 1.1


Link Measurements for the Router1 Subnetwork
bytes/sec
1000000
750000
500000
250000
0
corp
local
switch
fddi
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
00:00
hour of day
as.factor(dir)
dst
src

Question 1.2


Local Variances Versus Local Means for Router1 Link Measurements
time 11:30
time 15:30
log10(var)
log10(mean)
10
9
8
7
4.4
4.8
5.2


Local Variances Versus Local Means for Router2 Link Measurements
time 11:30
time 15:30
log10(var)
log10(mean)
12
10
8
6
4
3
4
5

Question 1.3

The unobserved OD byte counts, $\mathbf{x}_t$ at time t are modeled as a vector of independent normal random variables:

$$\mathbf{x}_t \sim N(\lambda, \mathbf{\Sigma}) \tag{1}$$

The complete data log-likelihood is:

$$l(\theta|\mathbf{X}) = -\frac{T}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\sum_{t=1}^{T}(\mathbf{x}_t - \lambda)'\mathbf{\Sigma}^{-1}(\mathbf{x}_t - \lambda) \tag{2}$$

The EM conditional expectation function Q is

$$\begin{aligned} Q(\theta, \theta^{(k)}) &= \mathbf{E}(l(\theta|\mathbf{X})|\mathbf{Y}, \theta^{(k)}) && (3)\\ &= \mathbf{E}\left[-\frac{T}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\sum_{t=1}^{T}(\mathbf{x}_t - \lambda)'\mathbf{\Sigma}^{-1}(\mathbf{x}_t - \lambda)\right] && (4)\\ &= -\frac{T}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\sum_{t=1}^{T}\mathbf{E}\left[(\mathbf{x}_t - \lambda)'\mathbf{\Sigma}^{-1}(\mathbf{x}_t - \lambda)\right] && (5)\end{aligned}$$

Since $(\mathbf{x}_t - \lambda)'\mathbf{\Sigma}^{-1}(\mathbf{x}_t - \lambda)$ has a quadratic form, we can use the formula for its expectation

$$\begin{aligned} \mathbf{E}\left[(\mathbf{x}_t - \lambda)'\mathbf{\Sigma}^{-1}(\mathbf{x}_t - \lambda)\right] &= \mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathrm{Var}(\mathbf{x}_t - \lambda|\mathbf{y}_t, \theta^{(k)})\right) + \mathbf{E}(\mathbf{x}_t - \lambda|\mathbf{y}_t, \theta^{(k)})'\mathbf{\Sigma}^{-1}\mathbf{E}(\mathbf{x}_t - \lambda|\mathbf{y}_t, \theta^{(k)}) && (6)\\ &= \mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathrm{Var}(\mathbf{x}_t|\mathbf{y}_t, \theta^{(k)})\right) + (\mathbf{m}_t^{(k)} - \lambda)'\mathbf{\Sigma}^{-1}(\mathbf{m}_t^{(k)} - \lambda) && (7)\\ &= \mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{R}^{(k)}\right) + (\mathbf{m}_t^{(k)} - \lambda)'\mathbf{\Sigma}^{-1}(\mathbf{m}_t^{(k)} - \lambda) && (8)\end{aligned}$$

where

$$\begin{aligned} \mathbf{m}_t^{(k)} &= \mathbf{E}(\mathbf{x}_t|\mathbf{y}_t, \theta^{(k)}) && (9)\\ &= \lambda^{(k)} + \mathbf{\Sigma}^{(k)}\mathbf{A}'(\mathbf{A}\mathbf{\Sigma}^{(k)}\mathbf{A}')^{-1}(\mathbf{y}_t - \mathbf{A}\lambda^{(k)}), && (10)\\ \mathbf{R}^{(k)} &= \mathrm{Var}(\mathbf{x}_t|\mathbf{y}_t, \theta^{(k)}) && (11)\\ &= \mathbf{\Sigma}^{(k)} - \mathbf{\Sigma}^{(k)}\mathbf{A}'(\mathbf{A}\mathbf{\Sigma}^{(k)}\mathbf{A}')^{-1}\mathbf{A}\mathbf{\Sigma}^{(k)} && (12)\end{aligned}$$

Here we have taken advantage of the conditional distributions of a multivariate normal.
We can plug (8) back into (5) to obtain eq. 6 in Cao et al. (2000), completing the derivation.

The EM algorithm

1. E-step: Calculate $\mathbf{m}_j^{(k)}$ and $\mathbf{R}^{(k)}$. If $k = 0$ we must initialize the values.
2. M-step: Maximize $Q(\theta, \theta^{(k)})$ with respect to θ.
3. Repeat steps 1 and 2 until convergence.

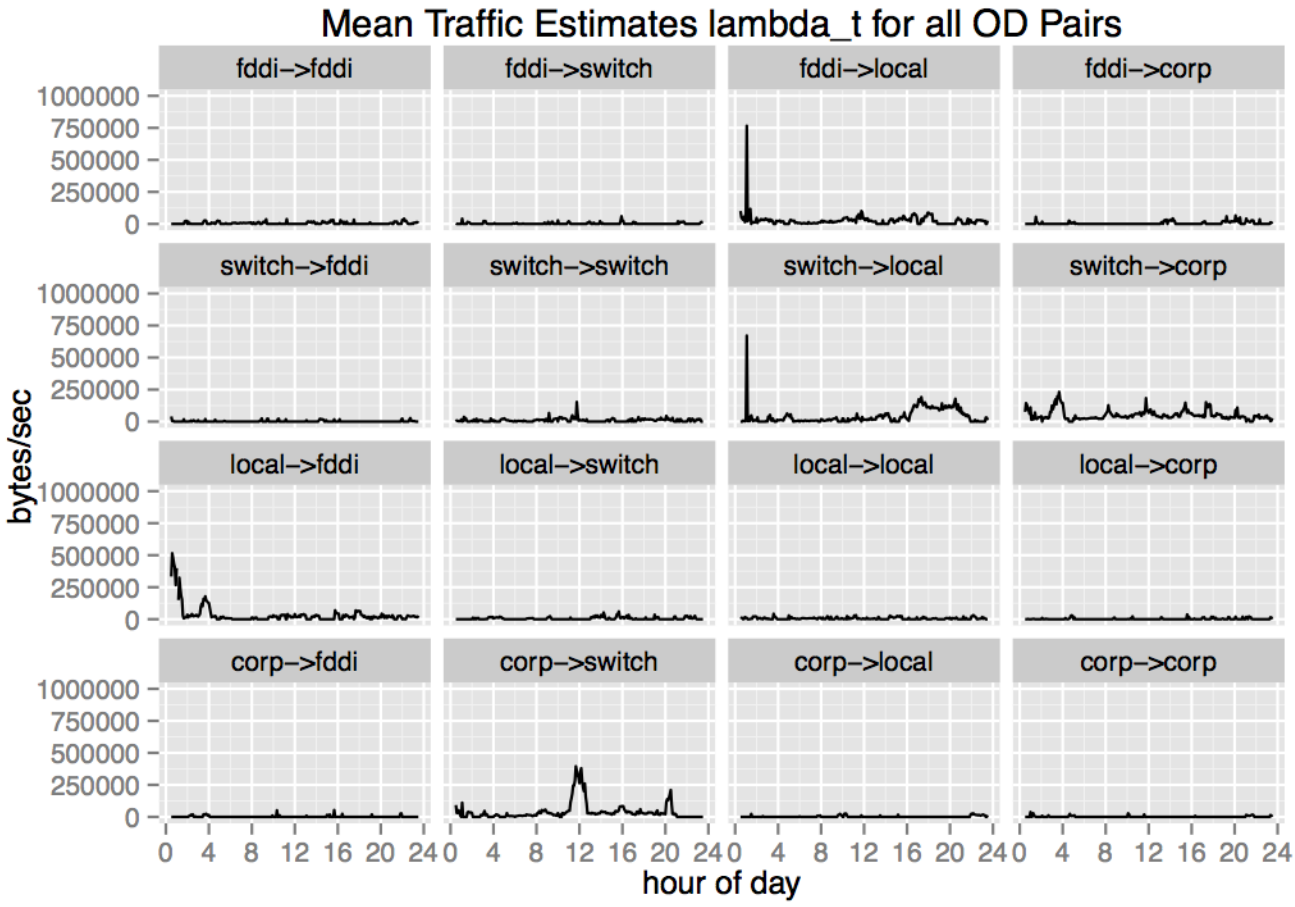
Question 1.4

To fit the EM algorithm derived for the iid model to the locally iid model, we simply modify the Q function. To estimate the parameters at each time point t, instead of summing over all time points $[1, T]$, we only sum

over time points in our window, which are assumed to be iid normal:

$$Q(\theta, \theta^{(k)}) = \mathbf{E}\left[-\frac{w}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\sum_{j=t-h}^{t+h}(\mathbf{x}_j - \lambda)'\mathbf{\Sigma}^{-1}(\mathbf{x}_j - \lambda)\right] \tag{13}$$

Initializing θ to $\vec{0}$ and running EM until an absolute convergence threshold of 1E-4, we successfully replicated Figure 5:



Question 1.5

In this section, we let $\eta_t = (\log(\lambda_t), \phi(t))$, and have the update $\eta_t = \eta_{t-1} + v_t$, where $v_t \sim N(0, V)$, where V is a fixed variance matrix that we choose beforehand (per suggestions, we set it equal to 5*diag(exp(lambda))).

We have that $p(\eta_t|\tilde{Y}_t) = p(\eta_t|\tilde{Y}_{t-1}, Y_t) \propto p(\eta_t|\tilde{Y}_{t-1})p(Y_t|\eta_t)$, so that $\log p(\eta_t|\tilde{Y}_{t-1}, Y_t) \propto \log p(\eta_t|\tilde{Y}_{t-1}) + \log p(Y_t|\eta_t)$. The first term on the right hand side is the prior and the second is the likelihood.

Thus, the process for maximizing our posterior $\log p(\eta_t|\tilde{Y}_{t-1}, Y_t)$ is the same as before except with our additive penalty correction $\log p(\eta_t|\tilde{Y}_{t-1})$.

We have that

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

If we approximate $p(\eta_{t-1}|\tilde{Y}_{t-1})$ by $N(\hat{\eta}_{t-1}, \hat{\Sigma}_{t-1})$, where $\hat{\eta}_{t-1}$ is the posterior mode and $\hat{\Sigma}_{t-1}$ is the inverse of the curvature of the log posterior density at the mode, then we can approximate $p(\eta_t|\tilde{Y}_{t-1})$ by $N(\hat{\eta}_{t-1}, \hat{\Sigma}_{t-1} + V)$.

So now our Q function is the same as that from part 4 except with this added term

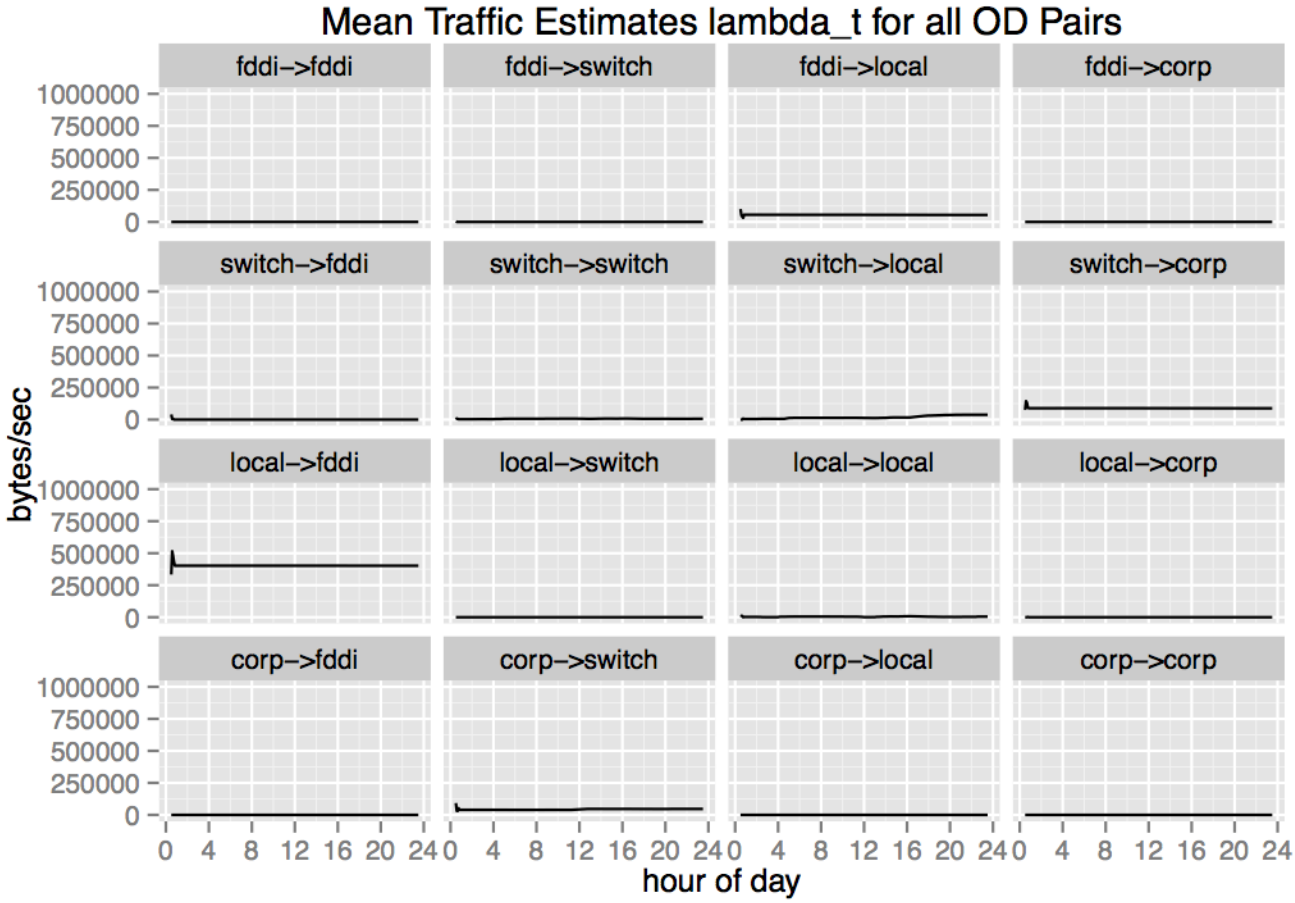
$$-\frac{w}{2}\log|\hat{\Sigma}_{t-1} + V| - \frac{1}{2}(\eta - \hat{\eta}^{(k)})' \left(\hat{\mathbf{\Sigma}}_{\mathbf{t-1}} + \mathbf{V}\right)^{-1} (\eta - \hat{\eta}^{(k)})$$

Our algorithm is then as described in the paper:

1. Initialize $t = h + 1, \hat{\eta}_{t-1} = \hat{\eta}_0, \Sigma_{t-1} = \Sigma_0$
2. Set $\hat{\Sigma}_{t|t-1} = \hat{\Sigma}_{t-1} + V$ and $\pi(\eta_t) = p(\eta_t|\tilde{Y}_{t-1}) \sim N(\hat{\eta}_{t-1}, \hat{\Sigma}_{t|t-1})$, which is the prior of η_t given observations $[scale = 1]YoungGentili_fig5.png$
   Y_{t-1} as described before.
3. Define $g(\eta_t)$ to be $\log \pi(\eta_t) + \log p(Y_t|\eta_t)$. Find the mode $\hat{\eta}_t$ using optim
4. Update $\hat{\Sigma}_t = \ddot{g}(\hat{\eta}_t)^{-1}$, where $\ddot{g}(\eta_t) = -\hat{\Sigma}^{-1}_{t|t-1} + \dfrac{\partial^2 \log p}{\partial \eta_t^2}$

Question 1.6

Though evidence of smoothing is shown, we had some issues reproducing Figure 6, in particular evaluating the Hessian matrix.



Question 1.8

The method of Cao et al is more adaptive than that of Tebaldi and West. More specifically, it allows for continuous measurements of data over time, and changes in the parameters of the model accordingly. Tebaldi and West instead fix their model for a certain time interval. Networks are dynamic (there may be more traffic during morning hours, after people wake up, and a lull after 5pm, when people leave work, for instance), so we except that a fixed model will not be able to capture all the nuances of time-varying traffic.

Tebaldi and West also assume that the network traffic is iid Poisson distributed, while Cao et al say that network traffic can be approximated by iid normals. Cao et al claim that such an approximation is valid given the high speed of today's networks and the use of 5 minute intervals, which makes the discreteness of byte counts negligible. Furthermore, by introducing a ϕ parameter, Cao and Yu are able to accommodate a change in units of traffic measurement, so their model is scale-invariant while that of Tebaldi is not. Another benefit of using iid normals rather than iid poissons is the computational efficiency of the former and the computational complexity of the latter. With the iid normal assumption, Cao et al are able to derive equations that allow for use of the Expectation Maximization algorithm to compute the optimal x vector. Tebaldi and West instead used MCMC to sample from the posterior distribution.

Cao et al's model also uses a power relation, which is an extension to the Poisson model that allows

for overdispersion and a superlinear increase in the variance with the mean, once again making their model more powerful than that of Tebaldi and West.

Question 1.9

Below is the locally iid model fit to the link loads of router 2: