10-708: Probabilistic Graphical Models 10-708, Spring 2017

Homework2

Lecturer: Eric P. Xing *Name: Yuan Liu(yuanl4)*

1 Problem 1 EM algorithm for Confounded Heterogeneous Data

1.1 EM algorithm

(a) What is the sufficient statistics for σ_μ?

$\sqrt{u^T u}$, we can find the reason of this answer in subproblem (c).

(b) Derive the MLE for σ_u, σ_ϵ and β.

Because $y = X\beta + Zu + \epsilon$ and $y \sim N(X\beta, K\sigma_u^2 + I\sigma_\epsilon^2)$, we can get $Cov(y, u) = E(y - X\beta)u^T = E(Zu + \epsilon)u^T = ZE(uu^T) = \sigma_u^2 Z$. Then we will get the following distribution:

$$\begin{bmatrix} y \\ u \end{bmatrix} \sim N(\begin{bmatrix} X\beta \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 K + I\sigma_\epsilon^2 & \sigma_u^2 Z \\ \sigma_u^2 Z^T & \sigma_u^2 I \end{bmatrix}) \tag{1}$$

Then we can calculate the conditional distribution:

$$E(y|u) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y - \mu_2) = X\beta + Zu$$

$$V(y|u) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \sigma_\epsilon^2 I$$

$$\Rightarrow y|u \sim N(X\beta + Zu, \sigma_\epsilon^2 I)$$

Then we can begin to calculate the likelihood.

$$\begin{aligned} \log p(y, u) &= \log p(y|u)p(u) = \log p(y|u) + \log p(u) \\ &\propto -\frac{1}{2}\log|\sigma_\epsilon^2 I| - \frac{1}{2}Tr((y - X\beta - Zu)(y - X\beta - Zu)^T(\sigma_\epsilon^2 I)^{-1}) \\ &\quad -\frac{1}{2}\log|\sigma_u^2 I| - \frac{1}{2}Tr(uu^T(\sigma_u^2 I)^{-1}) \\ &= -n\log\sigma_\epsilon - q\log\sigma_u - \frac{1}{2\sigma_\epsilon^2}(y - X\beta - Zu)^T(y - X\beta - Zu) - \frac{1}{2\sigma_u^2}u^T u \end{aligned}$$

Then by the derivation of the log likelihood function, we can find the expression of σ_u,σ_ϵ and β

$$\frac{\log p(y, u)}{\partial\sigma_u} = -\frac{q}{\sigma_u} + \frac{u^T u}{\sigma_u^3}$$

$$\Rightarrow \hat{\sigma}_u = \sqrt{\frac{u^T u}{q}}$$

$$\frac{\log p(y,u)}{\partial \sigma_\epsilon} = -\frac{n}{\sigma_\epsilon} + \frac{(y - X\beta - Zu)^T(y - X\beta - Zu)}{\sigma_\epsilon^3}$$
$$\Rightarrow \hat{\sigma}_\epsilon = \sqrt{\frac{(y - X\beta - Zu)^T(y - X\beta - Zu)}{n}}$$
$$\frac{\log p(y,u)}{\partial \beta} = \frac{1}{\sigma_\epsilon^2} X^T(y - X\beta - Zu)$$
$$\hat{\beta} = (X^T X)^{-1} X^T (y - Zu)$$

(c) Derive the EM algorithm to estimate σ_μ and β. According to (1), we can get following(The third formula has the same meaning of $E[X^2] = V(X) + E[X]^2$):

$$E(u|y) = \sigma_u^2 Z^T(\sigma_u^2 K + \sigma_\epsilon^2 I)^{-1}(y - X\beta)$$
$$V(u|y) = \sigma_u^2 I - \sigma_u^4 Z^T(\sigma_u^2 K + \sigma_\epsilon^2 I)^{-1} Z$$
$$E(u^T u|y) = Tr(V(u|y)) + ||E(u|y)||_2^2$$

So by these, we can write down our EM algorithm:

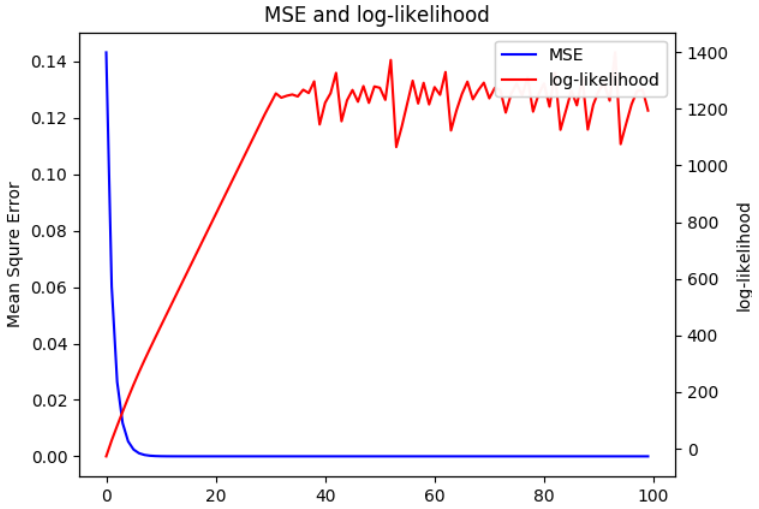
E Step:

$$\langle u \rangle = \sigma_u^2 Z^T(\sigma_u^2 K + \sigma_\epsilon^2 I)^{-1}(y - X\beta)$$
$$\langle u^T u \rangle = Tr(\sigma_u^2 I - \sigma_u^4 Z^T(\sigma_u^2 K + \sigma_\epsilon^2 I)^{-1} Z) + ||\sigma_u^2 Z^T(\sigma_u^2 K + \sigma_\epsilon^2 I)^{-1}(y - X\beta)||_2^2$$

M Step:

$$\sigma_u = \sqrt{\frac{\langle u^T u \rangle}{q}}$$
$$\sigma_\epsilon = \sqrt{\frac{(y - X\beta - Z\langle u \rangle)^T(y - X\beta - Z\langle u \rangle)}{n}}$$
$$\beta = (X^T X)^{-1} X^T (y - Z\langle u \rangle)$$

1.2 A Playground



The MSE and log-likelihood are calculated by the following two formulas in each iteration.

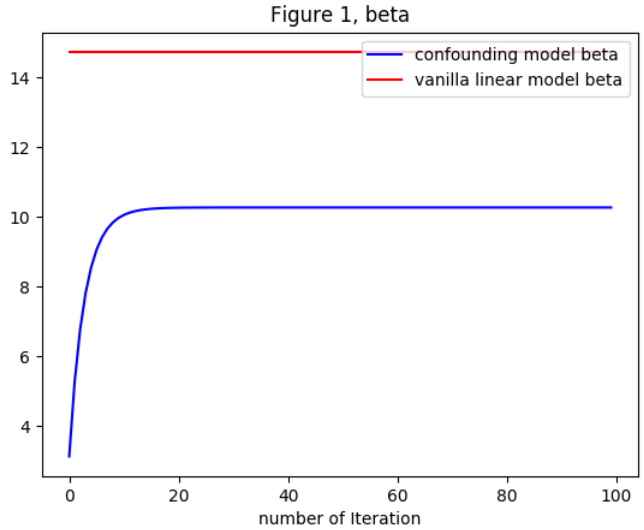
$$MSE = ||\beta - \beta_{golden}||_2^2$$

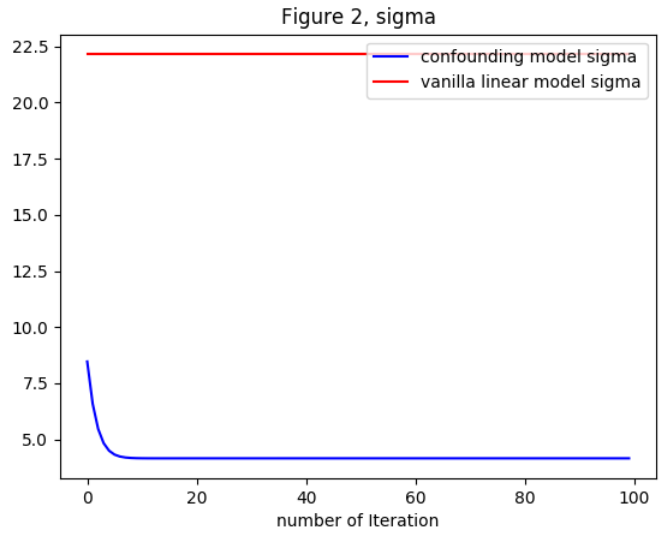
$$log - likelihood = -n \log \sigma_\epsilon - q \log \sigma_u - \frac{1}{2\sigma_\epsilon^2}(y - X\beta - Z\langle u \rangle)^T (y - X\beta - Z\langle u \rangle) - \frac{1}{2\sigma_u^2}\langle u^T u \rangle$$

1.3 Let's GO Wild

(a) The dataset used here can be found in the following site
https://vincentarelbundock.github.io/Rdatasets/datasets.html
The datasets is about the Effect of Vitamin C on Tooth Growth in Guinea Pigs. In this datasets, the response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice or ascorbic acid (a form of vitamin C and coded as VC). So we use y to present the length of odontoblasts, and use x to present the dose levels of vitamin C. We can assume vitamin C has some benefit to the length of odontoblasts, thus it will satisfy $E(y) = x\beta$. We use z to present the delivery methods. This two different methods may influence the efficiency of vitamin C absorption, thus it can influence the variance of $Var(y) = z^2\sigma_u^2 + \sigma_\epsilon^2$. So this dataset is suitable for this model.

(b) I draw the following two graphs to measure the confounding model and vanilla linear model. The first graph is about the value of β after each iteration. The second graph is about the value of σ_ϵ. We know if the model is good, the value of σ_ϵ should be small. According to figure 2, the confounding model performs better than the vanilla linear model, because σ_ϵ is smaller in the confounding model.



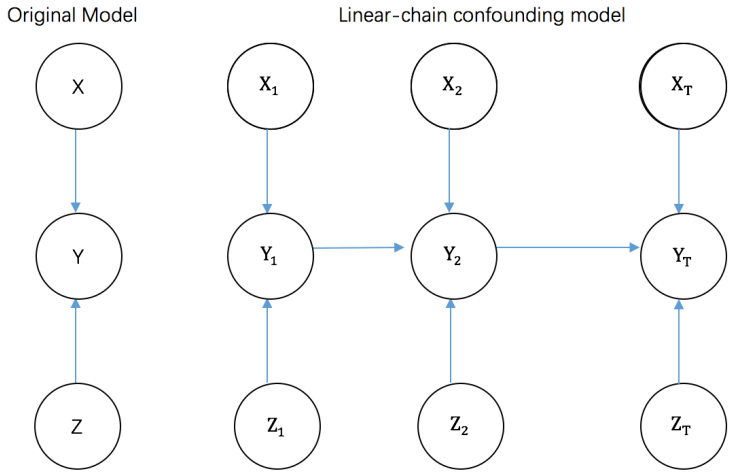


(c) The original model is showed in the following graph, by this model we can know

$$P(Y|X,Z) \sim N(X\beta, K\sigma_u^2 + I\sigma_\epsilon^2)$$

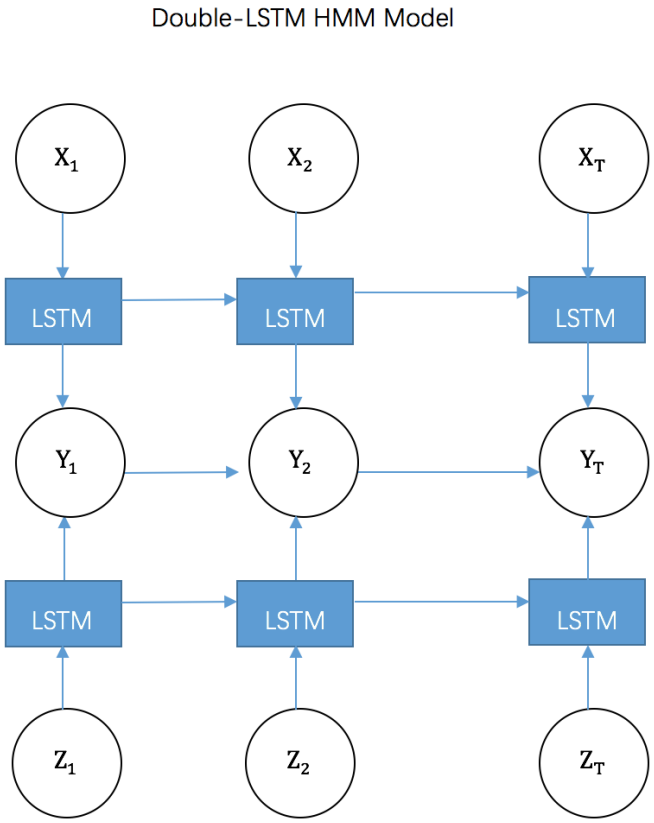
Then my first improvement is to make the original model to be a linear-chain confounding model. In this new model we need to add an transition matrix $P(Y_t|Y_{t-1})$.

$$P(\{Y_t\}|\{X\},\{Z_t\}) = P(Y_1|X_2,Z_2)\prod_{t=2}^{T} P(Y_t|X_t,Z_t)P(Y_t|Y_{t-1})$$



We can make this model even more wild, to use a LSTM layer to calculate $P(Y_t|X_t,Z_t)$. Then in this model we can calculate the $P(\{Y_t\}|\{X\},\{Z_t\})$ by the following formula. $P(Y_t|\{X_t\},\{Z_t\})$ is the output of LSTM layer.

$$P(\{Y_t\}|\{X\},\{Z_t\}) = P(Y_1|\{X_t\},\{Z_t\})\prod_{t=2}^{T} P(Y_t|\{X_t\},\{Z_t\})P(Y_t|Y_{t-1})$$


Double-LSTM HMM Model
X1
X2
XT
LSTM
LSTM
LSTM
Y1
Y2
YT
LSTM
LSTM
LSTM
Z1
Z2
ZT

2 Estimation of Precision Matrix

2.1 Precision Matrix

(a) By the definition of precision matrix P, the pdf of multi-normal distribution can be written in the following form(In this situation, we simply assume $E(X) = 0$):

$$p(x_1, x_2, ..., x_n) \propto \exp\{-\frac{\sum_i \sum_j P_{ij} x_i x_j}{2}\}$$

$$\begin{aligned} p(x_1, x_2|x_3, ..., x_n) &= \frac{p(x_1, x_2, ..., x_n)}{\int p(\tilde{x}_1, \tilde{x}_2, ..., x_n) d\tilde{x}_1 d\tilde{x}_2} \propto p(x_1, x_2, ..., x_n) \\ &\propto exp\{-\frac{P_{11}x_1^2 + 2P_{12}x_1x_2 + P_{22}x_2^2 + x_1(\sum_{i=3}^n P_{1,i}x_i) + x_2(\sum_{i=3}^n P_{2,i}x_i)}{2}\} \end{aligned}$$

We know $p(x_1, x_2|x_3, .., x_n)$ should in the form of 2-dimension normal distribution, which means:

$$p(x_1, x_2|x_3, ..., x_n) \propto \exp\{-\frac{1}{2(1-\rho^2)}[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}]\}$$

The second order term of these two representation should be equal, then we can get:

$$\frac{1}{(1-\rho^2)\sigma_1^2} = P_{11}$$

$$\frac{1}{(1-\rho^2)\sigma_2^2} = P_{22}$$

$$-\frac{\rho}{(1-\rho^2)\sigma_1\sigma_2} = P_{12}$$

Then we can get

$$\rho_{12} = -\frac{P_{12}}{\sqrt{P_{11}P_{22}}}$$

By simple extension, we can get the final relationship

$$\rho_{ij} = -\frac{P_{ij}}{\sqrt{P_{ii}P_{jj}}}$$

(b) By the similar way used in (a), we can get

$$p(x_1|x_2, ..., x_n) \propto exp\{-\frac{P_{11}x_1^2 + 2x_1(\sum_{i=2}^n P_{1,i}x_i)}{2}\}$$

By these we can get:

$$E(x_1|x_2, ..., x_n) = -\sum_{i=2}^n \frac{P_{1,i}}{P_{11}}x_i = \sum_{i=2}^n \rho_{1i}\sqrt{\frac{P_{ii}}{P_{11}}}x_i$$

$$\Rightarrow \sqrt{P_{11}}E(x_1|x_2, ..., x_n) = \sum_{i=2}^n \rho_{1i}\sqrt{P_{ii}}x_i$$

By simple extension, we can get:

$$\sqrt{P_{i,i}}E(x_i|x_{-i}) = \sum_{j\neq i} \rho_{i,j}\sqrt{P_{j,j}}x_j$$

(c) Assume $E(X) = 0$. We can get the log likelihood

$$l(\Sigma|\tilde{\Sigma}) \propto -\frac{1}{2}log|\Sigma| - \frac{1}{2}Tr(\tilde{\Sigma}\Sigma^{-1}) = \frac{1}{2}log|\Sigma^{-1}| - \frac{1}{2}Tr(\tilde{\Sigma}\Sigma^{-1})$$

So maximize log likelihood is equivalent to

$$\min -log|\Sigma^{-1}| + Tr(\tilde{\Sigma}\Sigma^{-1})$$

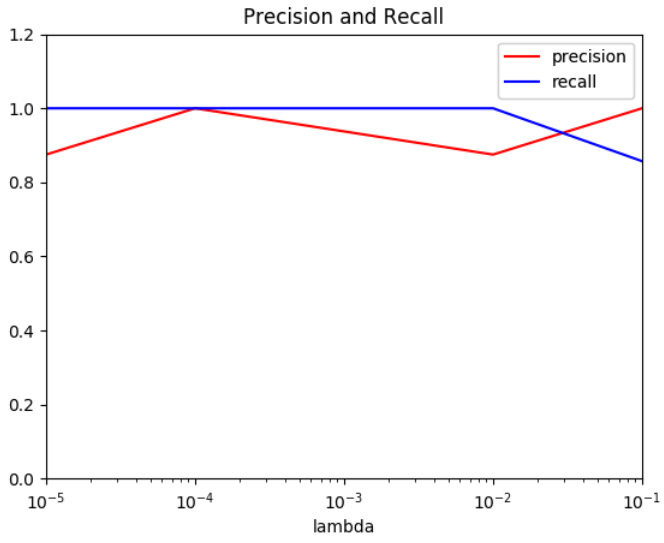
Let denote $P = \Sigma^{-1}$, then we can finally get

$$\hat{P} = arg\min_{P} -log|P| + Tr(\tilde{\Sigma}P)$$

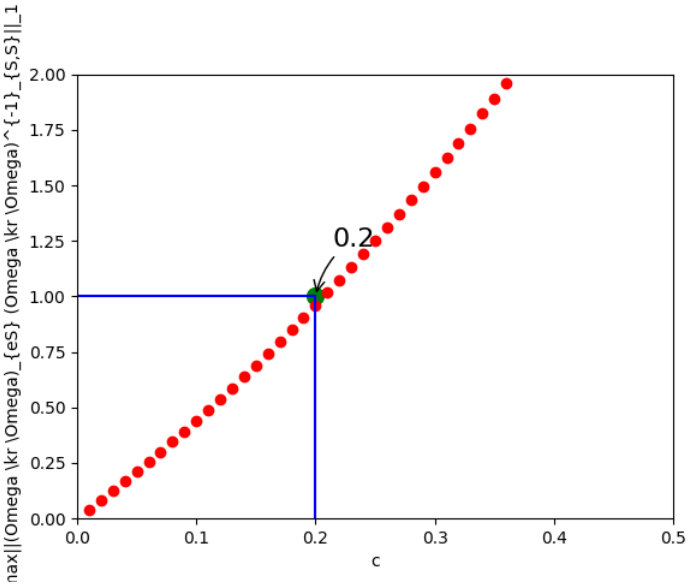
2.2 Graphical Lasso

(a) Implement Graphical Lasso

(b) By the code implemented in (a), we can draw the following picture.



(c) Calculate $\max_{e\in S^C} ||(\Sigma \otimes \Sigma)_{e,S}(\Sigma \otimes \Sigma)^{-1}_{S,S}||_1$ according to different c, we can get the following picture. Then we can know when $c \leq 0.2$, $\max_{e\in S^C} ||(\Sigma \otimes \Sigma)_{e,S}(\Sigma \otimes \Sigma)^{-1}_{S,S}||_1 < 1$. So we can get $c \in (0.2, 2^{-\frac{1}{2}}]$.



2.3 More about graphical Lasso

(a) Denote

$$L(P,\Sigma) \triangleq \frac{1}{2}tr(P^2\Sigma) - tr(P)$$

In order to prove this new cost function is convex, we can need to show two things:

- $\forall P_1, P_2, t \in (0,1)$, Denote $P \triangleq tP_1 + (1-t)P_2$, Then $tL(P_1,\Sigma) + (1-t)L(P_2,\Sigma) \geq L(P,\Sigma)$ Then we can calculate:

$$tL(P_1,\Sigma)+(1-t)L(P_2,\Sigma)-L(P,\Sigma) = \frac{t(1-t)}{2}tr((P_1-P_2)^2\Sigma) = \frac{t(1-t)}{2}tr((P_1-P_2)\Sigma(P_1-P_2))$$

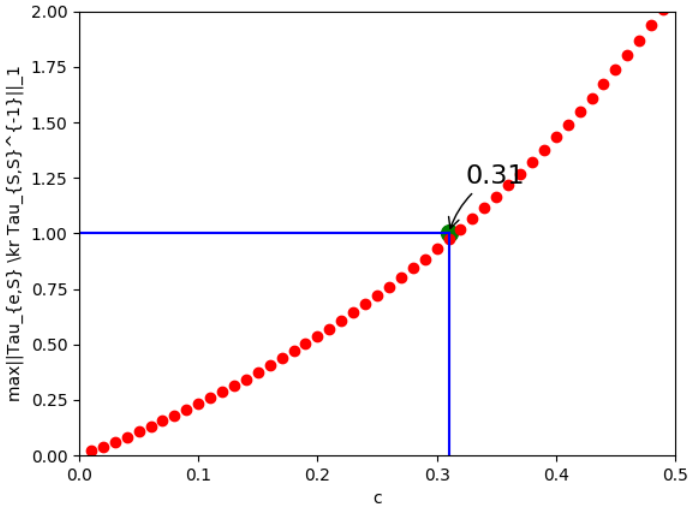
  Because we know Σ is positive definite, so the diagonal of $(P_1 - P_2)\Sigma(P_1 - P_2)$ should all be positive, then we can get $\frac{t(1-t)}{2}tr((P_1-P_2)\Sigma(P_1-P_2)) \geq 0$. By this we can proof the loss $L(P,\Sigma)$ is convex.

- The unique minimizer of $L(P,\Sigma)$ is Σ^{-1}. We can show it by the derivation of the loss

$$\frac{\partial L(P,\Sigma)}{\partial P} = \frac{1}{2}(\Sigma P + P\Sigma)^T - I$$

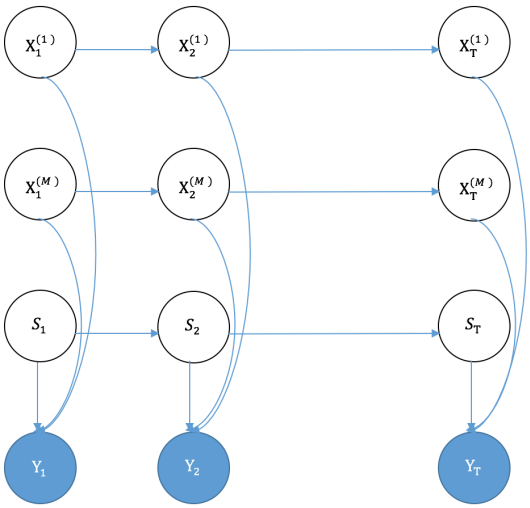
  By this we know when $P = \Sigma^{-1}$, the derivation $\frac{\partial L(P,\Sigma)}{\partial P} = 0$.

(b) Calculate $\max_{e\in S^C} ||\Gamma_{e,S}\Gamma^{-1}_{S,S}||_1$ according to different c, we can get the following picture. Then we can know when $c \leq 0.31$, $\max_{e\in S^C} ||\Gamma_{e,S}\Gamma^{-1}_{S,S}||_1 < 1$. So we can get $c \in (0.2, 0.31)$.


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
0.4
0.5
c
max||Tau_{e,S} \kr Tau_{S,S}^{-1}||_1
0.31

3 State Space Model

3.1 EM algorithm



Denote $\{Y_t\} = [Y_1, Y_2, ..., Y_T]$, $\{S_t, X_t, Y_t\}$, Then we can calculate the log-likelihood function

$$\begin{aligned}
\log P(\{Y_t\}|\theta) &= \log \sum_{\{S_t\}} \int P(\{S_t, X_t, Y_t\}|\theta) d\{X_t\} \\
&= \log \sum_{\{S_t\}} \int Q(\{S_t, X_t\}) [\frac{P(\{S_t, X_t, Y_t\}|\theta)}{Q(S_t, X_t)}] d\{X_t\} \\
&\geq \sum_{\{S_t\}} \int Q(\{S_t, X_t\}) log[\frac{P(\{S_t, X_t, Y_t\}|\theta)}{Q(S_t, X_t)}] d\{X_t\} \\
&\triangleq B(Q, \theta)
\end{aligned}$$

The E step is to find $Q^*(\theta) = argmin_Q B(Q, \theta)$. The solution of E-step is

$$Q(\{S_t, X_t\}) = P(\{S_t, X_t\}|\{Y_t\}, \theta)$$

We can view the switching state space model as a directed graphical model, which is showed in the figure. So $Q(\{S_t, X_t\}) = P(\{S_t, X_t\}|\{Y_t\}, \theta)$ involves the calculation of marginal distribution conditioned on $\{Y_t\}$. Then according to the model, we know that $S_t, Y_t, X_t^{(i)}$ and $X_t^{(i)}, Y_t, X_t^{(j)}$ forms v-structure. Because the observation of Y_t , $X_t^i|_{i=1}^M, S_t$ will become dependent. This induced dependency effectively couples all of the real-valued hidden state variables to the discrete switch variable, as a consequence of which the exact posteriors become Gaussian mixtures with an exponential number of terms.

3.2 Mean Field Method

We know

$$P(\{S_t, X_t, Y_t\}) = \left[P(S_1) \prod_{t=2}^{T} P(S_t|S_{t-1})\right] \prod_{m=1}^{M} \left[P(X_1^{(m)}) \prod_{t=2}^{T} P(X_t^{(m)}|X_{t-1}^{(m)})\right] \prod_{t=1}^{M} P(Y_t|X_t, S_t)$$

We use $\pi_i = P(S_1 = i)$ to denote the initial distribution, and $P(S_t = m|S_{t-1} = n) = \Phi^{(m,n)}$ to be the transition matrix, then we can get:

$$P(S_1) = \prod_{m=1}^{M} (\pi^{(m)})^{S_1^m}$$

$$P(S_t = m|S_{t-1} = n) = \prod_{m=1}^{M}\prod_{n=1}^{M} (\Phi^{(m,n)})^{S_t^{(m)} S_{t-1}^{(n)}}$$

$$P(Y_t|X_t, S_t) = \prod_{m=1}^{M} [P(Y_t|X_t, S_t = m)]^{S_t^{(m)}}$$

Then we want to use the following form to represent $P(\{S_t, X_t, Y_t\})$:

$$P(\{S_t, X_t, Y_t\}) = \frac{1}{Z} \exp\{-H(\{S_t, X_t, Y_t\})\}$$

$$\begin{aligned}
H = &\frac{1}{2}\sum_{m=1}^{M} log|Q^{(m)}| + \frac{1}{2}\sum_{m=1}^{M}(X_1^{(m)} - \mu_{X_1}^{(m)})'(Q^{(m)})^{(-1)}(X_1^{(m)} - \mu_{X_1}^{(m)}) \\
&+ \frac{T-1}{2}\sum_{m=1}^{M}\log|Q^{(m)}| + \frac{1}{2}\sum_{m=1}^{M}\sum_{t=2}^{T}(X_t^{(m)} - A^{(m)}X_{t-1}^{(m)})'(Q^{(m)})^{-1}(X_t^{(m)} - A^{(m)}X_{t-1}^{(m)}) \\
&+ \frac{T}{2}log|R| + \frac{1}{2}\sum_{m=1}^{M}\sum_{t=1}^{T} S_t^{(m)}(Y_t - C^{(m)}X_t^{(m)})'R^{-1}(Y_t - C^{(m)}X_t^{(m)}) \\
&- \sum_{m=1}^{M} S_1^{(m)}\log\pi^{(m)} - \sum_{t=2}^{T}\sum_{m=1}^{M}\sum_{n=1}^{M} S_t^{(m)}S_{t-1}^{(n)}\log\Phi^{(m,n)}
\end{aligned}$$

We can also rewrite ψ in the following form:

$$\psi(S_1) = P(S_1)q_1^{(S_1)} = \prod_{m=1}^{M}(\pi^{(m)}q_1^{(m)})^{S_1^{(m)}}$$

$$\psi(S_{t-1}, S_t) = P(S_t|S_{t-1})q_t^{(S_t)} = \prod_{m=1}^{M}\prod_{n=1}^{M}(\Phi^{(m,n)}q_t^{(m)})^{S_t^{(m)}S_{t-1}^{(n)}}$$

By the similar way we can get the H_Q of $Q(\{S_t, Y_t\})$.

$$\begin{aligned}
H_Q = &\frac{1}{2}\sum_{m=1}^{M} log|Q^{(m)}| + \frac{1}{2}\sum_{m=1}^{M}(X_1^{(m)} - \mu_{X_1}^{(m)})'(Q^{(m)})^{(-1)}(X_1^{(m)} - \mu_{X_1}^{(m)}) \\
&+ \frac{T-1}{2}\sum_{m=1}^{M}\log|Q^{(m)}| + \frac{1}{2}\sum_{m=1}^{M}\sum_{t=2}^{T}(X_t^{(m)} - A^{(m)}X_{t-1}^{(m)})'(Q^{(m)})^{-1}(X_t^{(m)} - A^{(m)}X_{t-1}^{(m)}) \\
&+ \frac{T}{2}log|R| + \frac{1}{2}\sum_{m=1}^{M}\sum_{t=1}^{T} h_t^{(m)}(Y_t - C^{(m)}X_t^{(m)})'R^{-1}(Y_t - C^{(m)}X_t^{(m)}) \\
&- \sum_{m=1}^{M} S_1^{(m)}\log\pi^{(m)} - \sum_{t=2}^{T}\sum_{m=1}^{M}\sum_{n=1}^{M} S_t^{(m)}S_{t-1}^{(n)}\log\Phi^{(m,n)} - \sum_{t=1}^{T}\sum_{m=1}^{M} S_t^{(m)}\log q_t^{(m)}
\end{aligned}$$

Then we can calculate the KL distance between $Q(\{S_t, Y_t\})$ and $P(\{S_t, X_t, Y_t\})$

$$\begin{aligned} KL(Q||P) &= \sum_{\{S_t\}} \int Q(\{S_t, X_t\}) log[\frac{P(\{S_t, X_t, Y_t\}|\theta)}{Q(S_t, X_t)}] d\{X_t\} \\ &= \langle H - H_Q \rangle_Q - \log Z_Q - logZ \end{aligned}$$

By the formula of H and H_Q, we can get

$$H_Q - H = \sum_{m=1}^{M} \sum_{t=1}^{T} \frac{1}{2} (h_t^{(m)} - S_t^{(m)})(Y_t - C^{(m)} X_t^{(m)})' R^{-1} (Y_t - C^{(m)} X_t^{(m)}) - S_t^{(m)} \log q_t^{(m)}$$

Then we can get

$$\frac{\partial \langle H - H_Q \rangle_Q}{\partial \langle S_t^{(m)} \rangle_Q} = -\log q_t^{(m)} - \frac{1}{2} \langle (Y_t - C^{(m)} X_t^{(m)})' R^{-1} (Y_t - C^{(m)} X_t^{(m)}) \rangle_Q$$

By setting $\frac{\partial \langle H - H_Q \rangle_Q}{\partial \langle S_t^{(m)} \rangle_Q} = 0$, we can get

$$q_t^{(m)} = \exp\{-\frac{1}{2} \langle (Y_t - C^{(m)} X_t^{(m)})' R^{-1} (Y_t - C^{(m)} X_t^{(m)}) \rangle\}$$

$$\frac{\partial \langle H - H_Q \rangle_Q}{\partial \langle X_t^{(m)} \rangle_Q} = -(h_t^{(m)} - \langle S_t^{(m)} \rangle_Q) \left((Y_t - C^{(m)} X_t^{(m)})' R^{-1} C^{(m)} \right)$$

By setting $\frac{\partial \langle H - H_Q \rangle_Q}{\partial \langle X_t^{(m)} \rangle_Q} = 0$, we can get

$$h_t^{(m)} = \langle S_t^{(m)} \rangle_Q = Q(S_t = m)$$