Additional File for Identifying Regulational Alterations in Gene Regulatory Networks by State Space Representation of Vector Autoregressive Models and Variational Annealing

Kaname Kojima\textsuperscript{1} Seiya Imoto\textsuperscript{1} Rui Yamaguchi\textsuperscript{1}
kame@ims.u-tokyo.ac.jp imoto@ims.u-tokyo.ac.jp rui@ims.u-tokyo.ac.jp
André Fujita\textsuperscript{2} Mai Yamauchi\textsuperscript{1} Noriko Gotoh\textsuperscript{1}
fujita@ime.usp.br cyowako@ims.u-tokyo.ac.jp ngotoh@ims.u-tokyo.ac.jp
Satoru Miyano\textsuperscript{1}
mnarmno@ims.u-tokyo.ac.jp

\textsuperscript{1}Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
\textsuperscript{2}Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010, São Paulo 05508-090, Brazil

1 Proof of Proposition 1 in Main Manuscript

We give a proof of Proposition 1 in the main manuscript.

**Proposition 1.** \( P(X, \Theta, E) \) is factorized into \( P(\Theta, E)P(X, E) \), and \( P(E_i|X, \Theta, E \backslash \{E_i\}) \) is given as a binomial distribution. Let \( \hat{Q}(E_i) \) be \( Q(E_i) \) maximizing the lower bound of the variational annealing for \( \tau \to +0 \) given by

\[
\int dE_1 \cdots dE_p \int dX \int d\Theta Q(X)Q(\Theta) \prod_i Q(E_i) \log \frac{P(X, \Theta, E)}{Q(X)Q(\Theta)\prod_i Q(E_i)^{\tau}}. \tag{1}
\]

Then, the set of \( E_i \in \{0,1\} \) maximizing \( \hat{Q}(E_i) \) is \( \arg\max_{E \in \{0,1\}^p} P(E) \).

**Proof.** We consider \( Q(E_i), Q(\Theta), \) and \( Q(X) \) maximizing Equation (1) for \( \tau \to +0 \). Since \( Q(E_i)^{\tau} \) is proportional to a binomial distribution, \( Q(E_i) \) converges to the Dirac delta function \( \delta(E_i, E_i) \) with some \( \bar{E}_i \in \{0,1\} \) for \( \tau \to +0 \). For \( Q(E_i) = \delta(E_i, \bar{E}_i \} \) and \( \tau \to +0 \), \( Q(\Theta) \) and \( Q(X) \) maximizing Equation (1) are respectively given as \( P(\Theta|\bar{E}) \) and \( P(X|\bar{E}) \), where \( \bar{E} \) is the set of \( E_i \).

On the other hand, for \( Q(\Theta) = P(\Theta|\bar{E}) \), \( Q(X) = P(X|\bar{E}) \), and \( \tau \to +0 \), \( Q(E_i) \) converges to \( \delta(E_i, E_i^*) \) with some \( E_i^* \in \{0,1\} \), and Equation (1) is given by

\[
\int dE_1 \cdots dE_p \int dX \int d\Theta \prod_i \delta(E_i, E_i^*) P(X|\bar{E}) P(\Theta|\bar{E}) \log \frac{P(X|\bar{E})P(\Theta|\bar{E})P(E)}{P(X|\bar{E})P(\Theta|\bar{E})} = \int P(X|\bar{E}) \log \frac{P(X|E^*)}{P(X|\bar{E})} dX + \int P(\Theta|\bar{E}) \log \frac{P(\Theta|E^*)}{P(\Theta|\bar{E})} d\Theta + \log P(E^*), \tag{2}
\]

where \( E^* \) is the set of \( E_i^* \). \( E_i^* \) maximizing Equation (2) is \( \hat{E}_i \), and for \( E_i^* = \hat{E}_i \), Equation (2) amounts to \( \log P(\bar{E}) \). Therefore, \( Q(\hat{E}_i), Q(\Theta), \) and \( Q(X) \) maximizing Equation (1) are respectively \( \delta(E_i, \hat{E}_i) \), \( P(\Theta|\bar{E}) \), and \( P(X|\bar{E}) \) with \( \bar{E} = \arg\max_{E \in \{0,1\}^p} P(E) \). \qed
2 More Details on Procedures of Variational Annealing on Proposed Model

In the variational annealing on the proposed model, we calculate $Q$ functions for hidden variables $X$, parameters $\Theta$, and binary variables $E$ iteratively while cooling temperature $\tau$ to zero gradually at each iteration cycle. Here, we show the calculation procedures of $Q(X)$, $Q(\Theta)$, and $Q(E)$ as variational E-step, variational M-step, and variational A-step, respectively under the complete likelihood of the proposed model:

$$P(Y, X, \Theta, E) = \prod_{c=1}^{2} \prod_{t \in T^{(c)}} \frac{|H|^{-1/2}}{2\pi^{|H|}} \exp \left\{ -\frac{1}{2} (x^{(c)}_t - A \circ E^{(c)} x^{(c)}_{t-1})' H^{-1} (x^{(c)}_t - A \circ E^{(c)} x^{(c)}_{t-1}) \right\}$$

$$\times \prod_{t \in T^{(c)}_{obs}} \frac{|R|^{-1/2}}{2\pi^{|R|}} \exp \left\{ -\frac{1}{2} (y^c_t - x^{(c)}_t)' R^{-1} (y^c_t - x^{(c)}_t) \right\} P(\Theta, E),$$

For the notational brevity, we denote the expectation of a value $x$ with a probability distribution $Q(y)$ as $\langle x \rangle_{Q(y)}$.

2.1 Variational E-step

The proposed model is considered as the state space model in terms of hidden variables $X = \{x^{(c)}_t\}$. In the state space model, system matrix is given by $A \circ E^{(c)}$, and observation matrix is a $p$-dimensional identity matrix. Therefore, the parameters of $Q(X)$ are mean of $x_t$, variance of $x_t$, and cross time variance of $x_{t-1}$ and $x_t$. These parameters can be calculated via variational Kalman filter by using following terms expected with $Q(\Theta)$ and $Q(E)$:

$$\langle E^{(c)}_t \rangle_{Q(\Theta)} Q(E), \langle A_t \rangle_{Q(\Theta)} Q(E), \langle H^{-1} A \circ E^{(c)} \rangle_{Q(\Theta)} Q(E),$$

and

$$\langle (A \circ E^{(c)})' H^{-1} A \circ E^{(c)} \rangle_{Q(\Theta)} Q(E).$$

For the details of variational Kalman filter, see Chapter 5 of [1]. Let mean of $x_t$, variance of $x_t$, and cross time variance of $x_{t-1}$ and $x_t$ be $\mu_{x_t}$, $\Sigma_t$, and $\Sigma_{t,t-1}$, respectively. From the parameters of $Q(X)$, expectations of $x^{(c)}_t$, $x^{(c)}_t \left( x^{(c)}_t \right)'$, and $x^{(c)}_{t+1} \left( x^{(c)}_{t+1} \right)'$ with $Q(X)$ required in other steps are calculated as follows:

$$\langle x^{(c)}_t \rangle_{Q(X)} = \mu_{x_t},$$

$$\langle x^{(c)}_t (x^{(c)}_t)' \rangle_{Q(X)} = \mu_{x_t} (\mu_{x_t})' + \Sigma_t,$$

$$\langle x^{(c)}_{t-1} (x^{(c)}_{t-1})' \rangle_{Q(X)} = \mu_{x_{t-1}} (\mu_{x_{t-1}})' + \Sigma_{t-1,t}.$$

2.2 Variational M-step

$Q(\Theta)$ is factorized into $\prod_i Q(A_i | h_i) Q(h_i) Q(r_i) \prod_j Q(z_j)$, where $A_i$ is a vector given by $(A_{i1}, \ldots, A_{ip})'$. From the design of the proposed model, $Q(A_i | h_i)$, $Q(h_i)$, $Q(r_i)$, and $Q(z_j)$ are given in the following form:

$$Q(A_i | h_i) = N(A_i; \mu_{A_i}, h_i T_{A_i}^{-1}),$$

$$Q(h_i) = I G(h_i; u_i, k_i),$$

$$Q(r_i) = I G(r_i; v_i, l_i),$$

$$Q(z_j) = B(\zeta_{i0}; \zeta_{i1}).$$

Here, $T_{A_i}$ is a matrix given by $\sum_{c=1}^{2} \sum_{t=1}^{T^{(c)}-1} \langle x^{(c)}_t \left( x^{(c)}_t \right)' \rangle Q(X) \circ \langle E^{(c)}_t (E^{(c)}_t)' \rangle Q(E)$, where $E^{(c)}_t$ is a vector given by $(E^{(c)}_{i1}, \ldots, E^{(c)}_{ip})'$, and $\mu_{A_i}$ is given by $T_{A_i}^{-1} \sum_{c=1}^{2} \sum_{t=1}^{T^{(c)}-1} \langle x^{(c)}_t \left( x^{(c)}_t \right)' \rangle Q(X) \langle E^{(c)}_t \rangle Q(E)$.
$u_i, k_i, v_i,$ and $l_i$ are given as follows:

$$
 u_i = u_0 + \frac{1}{2} \sum_{c=1}^{2} (T^{(c)} - 1),
$$

$$
 k_i = k_0 + \frac{1}{2} \sum_{c=1}^{2} T^{(c)}
$$

$$
 v_i = v_0 + \frac{1}{2} \sum_{c=1}^{2} |T^{(c)}|,
$$

$$
 l_i = l_0 + \frac{1}{2} \sum_{c=1}^{2} \sum_{t \in T^{(c)}} \left( (y_{t,i}^{(c)})^2 - 2y_{t,i}^{(c)} \langle x_{t,i}^{(c)} \rangle_{Q(X)} + \langle (x_{t,i}^{(c)})^2 \rangle_{Q(X)} \right),
$$

where $x_{t,i}^{(c)}$ and $y_{t,i}^{(c)}$ are the $i$th element of $x_{i}^{(c)}$ and $y_{i}^{(c)}$, respectively.

$$
 \zeta_{i,0} = \zeta_0 + (E_{ij}^{(1)})\langle E_{ij}^{(2)} \rangle_{Q(E)} + (1 - (E_{ij}^{(1)})\langle E_{ij}^{(2)} \rangle_{Q(E)})\langle 1 - \langle E_{ij}^{(2)} \rangle_{Q(E)} \rangle
$$

$$
 \zeta_{i,1} = \zeta_1 + (1 - (E_{ij}^{(1)})\langle E_{ij}^{(2)} \rangle_{Q(E)})\langle E_{ij}^{(1)} \rangle_{Q(E)} + \langle E_{ij}^{(1)} \rangle_{Q(E)}\langle 1 - \langle E_{ij}^{(2)} \rangle_{Q(E)} \rangle\langle 1 - \langle E_{ij}^{(2)} \rangle_{Q(E)} \rangle.
$$

By using the parameters, we consider the following expectations required for the calculation of variational E-step: $\langle E^{(c)} \rangle_{Q(\Theta)Q(E)}$, $\langle A \rangle_{Q(\Theta)Q(E)}$, $\langle H^{-1}A \circ E^{(c)} \rangle_{Q(\Theta)Q(E)}$, and $\langle (A \circ E^{(c)})' H^{-1}A \circ E^{(c)} \rangle_{Q(\Theta)Q(E)}$. $\langle E^{(c)} \rangle_{Q(E)}$ is given in the procedure of variational A-step, $\langle A \rangle_{Q(\Theta)}$ is given by $(\mu_A, \ldots, \mu_A)$, and $\langle H^{-1}A \circ E^{(c)} \rangle_{Q(\Theta)Q(E)}$ is given by $\langle (A \circ E^{(c)})' H^{-1}A \circ E^{(c)} \rangle_{Q(\Theta)Q(E)}= \sum_i \langle (1/h_i)Q(h_i)\mu_A, \mu_A + T_A \rangle \circ \langle E^{(c)} \rangle_{Q(E)}$.

### 2.3 Variational A-step

For the calculation of $Q(E)$, we assume the factorization of $Q(E)$ to $\prod_c \prod_{ij} Q(E^{(c)}_{ij})$ in order to make the computation tractable. The likelihood with respect to $E_{ij}^{(c)}$ forms a binomial distribution. Note that $(E_{ij}^{(c)})^2$ is considered as $E_{ij}^{(c)}$ because $(E_{ij}^{(c)})^2 = E_{ij}^{(c)}$ holds for $E_{ij}^{(c)} = 0$ or 1. Thus, $Q(E^{(c)}_{ij})$ is given by a binomial distribution that takes one with probability $e_{ij}^{(c)}$ and 0 with probability $1 - e_{ij}^{(c)}$, and hence the expectation of $E_{ij}^{(c)}$ on $Q(E)$ is give as $e_{ij}^{(c)}$. For the preparation, we calculate $\langle A \rangle_{Q(\Theta)}$, $\langle H^{-1}A \rangle_{Q(\Theta)}$, $\langle A' H^{-1}A \rangle_{Q(\Theta)}$, $\langle x^{(c)}_i \rangle_{Q(X)}$, $\langle x^{(c)}_i (x^{(c)}_i)' \rangle_{Q(X)}$, and $\langle x^{(c)}_{i+1} (x^{(c)}_i)' \rangle_{Q(X)}$. $Q(E^{(c)}_{ij})$ is then iteratively calculated by using these expectations as well as the expectations $E_{ij}^{(c)}$ for $k \neq j$ on $Q(E)$. A few iterations are enough for the convergence. Let $w_{ij}^{(c)}$ be the $i$th element of $\sum_{t=1}^{T^{(c)-1}} \langle x^{(c)}_{t+1} (x^{(c)}_i)' \rangle_{Q(X)} \mu_A$, and $M_{ijk}$ be the $(j,k)$th element of

$$
 \sum_{t=1}^{T^{(c)-1}} \langle x^{(c)}_i (x^{(c)}_i)' \rangle_{Q(X)} \mu_A + T_A.
$$
Without loss of generality, we consider the calculation of $e^{(c)}_{ij}$ for $c = 1$. $e^{(1)}_{ij}$ is given by $\frac{\exp(\frac{1}{2}d_{ij}^{(1)})}{1+\exp(\frac{1}{2}d_{ij}^{(1)})}$, where

$$d_{ij}^{(1)} = w_{ij}^{(1)} + \frac{1}{2}M_{ij}^{(1)} + \sum_{j\neq i} E_{ik}^{(1)}M_{jk}^{(1)} + \log \sum_{E_{ij}^{(2)} = 0} \phi(E_{ij}^{(1)}, E_{ij}^{(2)})N(0, \alpha_1)F_{ij}N(0, \alpha_0)^{1-F_{ij}}.$$ 

By using the obtained $Q(E)$, we consider the expectations $\langle E_{ij}^{(c)} \rangle_{Q(E)}$ and $\langle E_{ij}^{(c)}(E_{ij}^{(c)})' \rangle_{Q(E)}$ required in the calculations of variational-E step and M-step. $\langle E_{ij}^{(c)} \rangle_{Q(E)}$ is given by $e^{(c)}_{ij}$, and the $(j, k)$th element of $\langle E_{ij}^{(c)}(E_{ij}^{(c)})' \rangle_{Q(E)}$ is given by $e^{(c)}_{ij}$ if $j = k$ holds and $e^{(c)}_{ij}e_{ik}^{(c)}$ otherwise.

3 More Details on Update of Hyperparameters

We show more details of updating hyperparams $u_0$, $k_0$, $v_0$, $l_0$, $\zeta_0$, and $\zeta_1$ by using the Newton-Raphson method. $u_0$ and $k_0$ are updated by maximizing the following equation:

$$\hat{u}_0(\hat{k}_0) = \arg \max_{(u_0, k_0)} \sum_i \int Q(h_i) \log IG(h_i; u_0, k_0)dh_i$$

$$= \arg \max_{(u_0, k_0)} (u_0 - 1) \frac{\sum_i (\log h_i)Q(h_i)}{p} + u_0 \log k_0 - k_0 \frac{\sum_i (1/h_i)Q(h_i)}{p} - \log \Gamma(u_0),$$

where $\langle \log h_i \rangle_{Q(h_i)}$ is given by $\log(k_i) - \psi(u_i)$ with the digamma function $\psi(\cdot)$. $\hat{u}_0$ and $\hat{k}_0$ are obtained by the Newton-Raphson method. Let $\gamma$ be a vector $(u_0, k_0)'$ and $f(\gamma)$ a function given by

$$f(\gamma) = (u_0 - 1) \frac{\sum_i (\log h_i)Q(h_i)}{p} + u_0 \log k_0 - k_0 \frac{\sum_i (1/h_i)Q(h_i)}{p} - \log \Gamma(u_0).$$

In the Newton-Raphson method, $\hat{u}_0$ and $\hat{k}_0$ are obtained by iteratively updating the following function:

$$\gamma^{s+1} = \gamma^s - \left( \frac{\partial^2 f(\gamma)}{\partial \gamma \partial \gamma'} \bigg|_{\gamma = \gamma^s} \right)^{-1} \frac{\partial f(\gamma)}{\partial \gamma} \bigg|_{\gamma = \gamma^s}. $$

Gradient and Hessian matrix of $f(\gamma)$ are given by

$$\frac{\partial f(\gamma)}{\partial \gamma} = \begin{pmatrix} -\psi(u_0) + \log k_0 + \frac{\sum_j (\log(1-z_{ij}))q(z_{ij})}{p} & u_0/k_0 - \frac{\sum_i (1/h_i)Q(h_i)}{p} \end{pmatrix},$$

and

$$\frac{\partial^2 f(\gamma)}{\partial \gamma \partial \gamma'} = \begin{pmatrix} -\psi_1(u_0) & 1/k_0 \\ 1/k_0 & -u_0/k_0^2 \end{pmatrix},$$

where $\psi_1(\cdot)$ is the trigamma function. $v_0$ and $l_0$ are also updated in a similar manner to $u_0$ and $k_0$. $\zeta_0$ and $\zeta_1$ are updated by solving the following equation:

$$\hat{\zeta}_0(\hat{\zeta}_1) = \arg \max_{\zeta_0, \zeta_1} \sum_j \int Q(z_{ij}) \log B(z_{ij}; \zeta_0, \zeta_1)dz_{ij}$$

$$= \arg \max_{\zeta_0, \zeta_1} (\zeta_1 - 1) \frac{\sum_j (\log(1-z_{ij}))Q(z_{ij})}{p} + (\zeta_0 - 1) \frac{\sum_j \log z_{ij}Q(z_{ij})}{p} - \log B(\zeta_0, \zeta_1).$$
Let $\zeta$ be a vector $(\zeta_0, \zeta_1)'$ and $g(\zeta)$ a function given by

$$g(\zeta) = (\zeta_1 - 1) \frac{\sum_j (\log(1 - z_{ij})) Q(z_{ij})}{p} + (\zeta_0 - 1) \frac{\sum_j (\log z_{ij}) Q(z_{ij})}{p} - \log B(\zeta_0, \zeta_1).$$

In the Newton-Raphson method, $\hat{\zeta}_0$ and $\hat{\zeta}_1$ are obtained by iteratively updating the following function:

$$\zeta^{s+1} = \zeta^s - \left( \frac{\partial^2 g(\zeta)}{\partial \zeta \partial \zeta'} \bigg|_{\zeta=\zeta^s} \right)^{-1} \left. \frac{\partial g(\zeta)}{\partial \zeta} \right|_{\zeta=\zeta^s},$$

where gradient and Hessian matrix of $g(\zeta)$ are given by

$$\frac{\partial g(\zeta)}{\partial \zeta} = \begin{pmatrix} \psi(\zeta_0 + \zeta_1) - \psi(\zeta_0) + \sum_j (\log z_{ij}) Q(z_{ij}) \\ \psi(\zeta_0 + \zeta_1) - \psi(\zeta_1) + \sum_j (\log(1 - z_{ij})) Q(z_{ij}) \end{pmatrix},$$

and

$$\frac{\partial^2 g(\zeta)}{\partial \zeta \partial \zeta'} = \begin{pmatrix} \psi_1(\zeta_0 + \zeta_1) - \psi_1(\zeta_0) & \psi_1(\zeta_0 + \zeta_1) \\ \psi_1(\zeta_0 + \zeta_1) & \psi_1(\zeta_0 + \zeta_1) - \psi_1(\zeta_1) \end{pmatrix}.$$

References

[1] Beal, M.J. (2003) Variational Algorithms for Approximate Bayesian Inference, PhD Thesis, Gatsby Computational Neuroscience Unit, University College London, London, UK.