Journal of Machine Learning Research 25 (2024) 1-55 Submitted 5/22; Revised 8/24; Published 9/24

Structured Optimal Variational Inference for Dynamic Latent Space Models

Peng Zhao PZHAO@UDEL.EDU
Department of Applied Economics and Statistics
University of Delaware
Newark, DE 19716, USA

Anirban Bhattacharya ANIRBANB@STAT.TAMU.EDU

Debdeep Pati DEBDEEP@STAT.TAMU.EDU

Bani K. Mallick BMALLICK@STAT.TAMU.EDU
Department of Statistics
Texas A&M University
College Station, TX 77843, USA

Editor: Ji Zhu

Abstract

We consider a latent space model for dynamic networks, where our objective is to estimate the pairwise inner products plus the intercept of the latent positions. To balance posterior inference and computational scalability, we consider a structured mean-field variational inference framework, where the time-dependent properties of the dynamic networks are exploited to facilitate computation and inference. Additionally, an easy-to-implement block coordinate ascent algorithm is developed with message-passing type updates in each block, whereas the complexity per iteration is linear with the number of nodes and time points. To certify the optimality, we demonstrate that the variational risk of the proposed variational inference approach attains the minimax optimal rate with only a logarithm factor under certain conditions. To this end, we first derive the minimax lower bound, which might be of independent interest. In addition, we show that the posterior under commonly adopted Gaussian random walk priors can achieve the minimax lower bound with only a logarithm factor. To the best of our knowledge, this is the first such a throughout theoretical analysis of Bayesian dynamic latent space models. Simulations and real data analysis demonstrate the efficacy of our methodology and the efficiency of our algorithm.

Keywords: Variational inference, dynamic network, hierarchical models, message-passing, posterior concentration

1. Introduction

Statistical analysis of network-valued data is rapidly gaining popularity in modern scientific research, with applications in diverse domains such as social, biological, and computer sciences to name a few. While there is now established literature on static networks (see, e.g., the survey articles by Goldenberg et al., 2010, Snijders, 2011 and Newman, 2018), the literature studying dynamic networks, that is, networks evolving over time, continues to show rapid growth; see Xing et al. (2010); Yang et al. (2011); Xu and Hero (2014); Hoff

©2024 Peng Zhao, Anirban Bhattacharya, Debdeep Pati and Bani K. Mallick.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/22-0514.html.

(2015); Sewell and Chen (2015); Matias and Miele (2017); Durante et al. (2017a); Durante and Dunson (2018); Pensky (2019) for a flavor.

The latent class model proposed in Hoff et al. (2002); see also Handcock et al. (2007); Hoff (2008); Krivitsky et al. (2009); Ma et al. (2020); constitutes an important class of static network models and has been widely used in visualization (Sewell and Chen, 2015), edge prediction (Durante et al., 2017b) and clustering (Ma et al., 2020). Latent space models represent each node i by a latent Euclidean vector $\boldsymbol{x}_i$, with the likelihood of an edge Y_{ij} between nodes i and j entirely characterized through some distance or discrepancy $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ between the respective latent coordinates. Dynamic extensions of latent space models (Sarkar and Moore, 2005; Sewell and Chen, 2015; Friel et al., 2016; Sewell and Chen, 2017; Liu and Chen, 2022; Loyal and Chen, 2023) are also available, which assume a Markovian evolution of the latent positions. We focus on statistical and computational aspects of variational inference in such dynamic latent space models in this article.

To set some preliminary notation, consider a network of n individuals observed over T time points. For $1 \leq i \neq j \leq n$, let Y_{ijt} denote the observed data corresponding to an edge between nodes i and j at time t. For example, $Y_{ijt} \in \{0, 1\}$ may denote the absence/presence of an edge, or $Y_{ijt} \in \mathbb{R}$ could indicate a measure of association between nodes i and j. Let $\boldsymbol{Y}_t = (Y_{ijt}) \in \mathbb{R}^{n \times n}$ denote the $n \times n$ network matrix at time t (with only the off-diagonal part relevant), and let $\mathcal{Y} = \{\boldsymbol{Y}_t\}_{t=1}^T$ denote the observed data. We formulate our latent space model using the commonly used negative inner product $d(\boldsymbol{x}_{it}, \boldsymbol{x}_{jt}) = -\boldsymbol{x}_{it}'\boldsymbol{x}_{jt}$ as the discrepancy measure (Durante et al., 2017b; Ma et al., 2020), where $\boldsymbol{x}_{it} \in \mathbb{R}^d$ denotes the latent Euclidean position of node i at time t and $\boldsymbol{x}'$ denotes the transpose of a vector $\boldsymbol{x}$. The observed data likelihood then takes the form

$$P(\mathcal{Y} \mid \mathcal{X}, \beta) = \prod_{t=1}^{T} \prod_{1 \leq i \neq j \leq n} P(Y_{ijt} \mid \beta, \boldsymbol{x}_{it}, \boldsymbol{x}_{jt}), \tag{1}$$

where $P(Y_{ijt} \mid \beta, \boldsymbol{x}_{it}, \boldsymbol{x}_{jt})$ is decided by $\beta + \boldsymbol{x}_{it}'\boldsymbol{x}_{jt}$, and $\mathcal{X} = \{\boldsymbol{X}_t\}_{t=1}^T$, with $\boldsymbol{X}_t = [\boldsymbol{x}_{1t}, ..., \boldsymbol{x}_{nt}]' \in \mathbb{R}^{n \times d}$ the matrix of the latent positions at time t and d is defined as the dimension of the latent space. To model the evolution of the latent positions, assume a Markov process

$$\begin{aligned} \boldsymbol{x}_{i1} &\sim \mathcal{N}(\boldsymbol{0}, \sigma_0^2 \mathbb{I}_d), \quad i = 1, ..., n. \\ \boldsymbol{x}_{i(t+1)} \mid \boldsymbol{x}_{it} &\sim \mathcal{N}(\boldsymbol{x}_{it}, \tau^2 \mathbb{I}_d), \quad i = 1, ..., n,\ t = 1, ..., T-1, \end{aligned} \tag{2}$$

where $\mathbb{I}_d$ is a $d \times d$ identity matrix.

To alleviate computational inefficiencies of sampling-based posterior inference, posterior approximations based variational inference have been developed where the variational posteriors of latent positions across all times are either directly (Liu and Chen, 2022) or implicitly (Sewell and Chen, 2017) assumed to be independent. In the paper by Sewell and Chen (2017), the variational family is presented in a joint form of $q(\mathcal{X})$ in their parametrization. However, in Section 2.2 of the supplementary document, the derivation only needs a fully factorized structure of the latent positions, and their algorithm for variational inference only obtains the marginal distributions $q(\boldsymbol{x}_{it})$ instead of joint ones. In dynamic models, where there is already *a priori* dependence between the latent states over time, assuming such an independent structure is restrictive and can lead to inconsistent estimation (Wang and Titterington, 2004).

In this article, we consider a more flexible *structured* mean-field (SMF) variational family, which only assumes a nodewise factorization. An efficient block coordinate ascent algorithm targeting the optimal SMF solution is developed, which scales linearly in the network size and retains the $O(nT)$ per-iteration computational cost of mean field (MF) by carefully constructing message-passing (MP) updates within each block to exploit the specific nature of the temporal dependence. Moreover, we empirically demonstrate that our algorithm achieves faster convergence across a wide range of simulated and real data examples. We also exhibit the mean of the optimal SMF solution to retain the same convergence rate as the exact posterior mean, providing strong support for its statistical accuracy. Overall, SMF achieves an optimal balance between the statistical accuracy of the exact posterior and the computational convenience of MF, retaining the best of both worlds. Loyal and Chen (2023) developed a similar SMF variational approach, and a coordinate ascent variational inference (CAVI) algorithm was introduced for a latent space model aimed at dynamic multilayer networks. Their study emphasized the algorithmic and computational perspectives, while one of our key emphasis is on deriving theoretical risk bounds for the proposed variational inference method below.

To adaptively learn the initial and transition standard derivations, we adopt priors

$$\sigma_0^2 \sim \text{Inverse-Gamma}\left(a_{\sigma_0}, b_{\sigma_0}\right), \quad \tau^2 \sim \text{Gamma}\left(c_\tau, d_\tau\right), \tag{3}$$

and incorporate them into our SMF framework. Although an inverse-gamma prior on the transition variance τ^2 (e.g., Sewell and Chen, 2015) leads to simple conjugate updates, it is now well-documented that an inverse-gamma prior on a lower-level variance parameter in Bayesian hierarchical models has undesirable properties when a strong shrinkage effect towards the prior mean is desired (Gelman, 2006; Gustafson et al., 2006; Polson and Scott, 2012). In contrast, adopting a Gamma prior (3) on τ^2 places sufficient mass near the origin, which aids our subsequent theoretical analysis and also retains closed-form updates in the form of Generalized inverse Gaussian distributions (Jorgensen, 2012).

From a theoretical perspective, statistical analysis of variational posteriors has received major attention recently (Pati et al., 2018; Wang and Blei, 2019; Alquier and Ridgway, 2020; Yang et al., 2020; Zhang and Gao, 2020). In particular, motivated by the recent development of Bayesian oracle inequalities for α-Rényi divergence risks (Bhattacharya et al., 2019), Alquier and Ridgway (2020) and Yang et al. (2020) proposed a theoretical framework, named α-Variational Bayes (α-VB), to analyze the variational risk of tempered or fractional posteriors in terms of α-Rényi divergences. Under the α-VB framework, statistical optimality of variational estimators can be guaranteed by sufficient prior concentration around the true parameter and appropriate control on the Kullback–Leibler (KL) divergence between a specific variational distribution and the prior. We adopt and extend their framework to derive Bayes risk bounds under the variational posterior towards the recovery of the latent positions in an appropriate metric. A novel ingredient of our theory is the ability to provide statistical analysis for SMF variational family $q(\mathcal{X}, \tau, \sigma_0) = q(\mathcal{X})q(\tau)q(\sigma_0)$ given hierarchically specified prior distributions of the form $p(\mathcal{X}, \tau, \sigma_0) = p(\mathcal{X} \mid \tau, \sigma_0)p(\tau)p(\sigma_0)$.

The proof technique of Alquier and Ridgway (2020) and Yang et al. (2020), where a specific variational candidate is constructed by truncating the prior to a small neighborhood around the true parameters, has become common for providing statistical guarantees of variational estimates. However, this technique cannot be directly applied to MF variational

families endowed with a hierarchical prior specification, as the truncated distribution may not be a candidate in MF variational family due to the dependence through the global prior in the hierarchy. Previous literature (Liu and Chen, 2022) avoids this issue by treating the upper-level parameters of the hierarchical prior as fixed constants, thus losing the adaptivity. Bai et al. (2020) developed statistical guarantees for full MF variational distribution in the context of regression with global local hierarchical priors. While their algorithm used a full MF family, their theoretical results are proven assuming a dependence between the upper and lower-level parameters of the hierarchy leading to a richer family rather than a fully factorized MF. On the other hand, our theoretical results and algorithm are both developed using the same structured mean-field family where the global parameter in the hierarchy is assumed to be independent of the remaining parameters.

In addition, we exhibit the optimality of our proposed variational estimator by showing its rate of convergence to be optimal up to a logarithmic term. En route, we identify an appropriate parameter space for the latent positions and derive information-theoretic lower bounds. To the best of our knowledge, this is the first derivation of a minimax lower bound for dynamic latent space models. In fact, the only other work we are aware of that studies minimax rates for dynamic network models is Pensky (2019) in the context of dynamic stochastic block models.

Finally, the computational and theoretical framework developed here can be safely adapted to the case where different nodes are equipped with different initials and transitions to capture nodewise differences:

$$\begin{aligned} &\boldsymbol{x}_{i1} \sim \mathcal{N}(\mathbf{0}, \sigma_{0i}^2 \mathbb{I}_d), && \boldsymbol{x}_{i(t+1)} \mid \boldsymbol{x}_{it} \sim \mathcal{N}(\boldsymbol{x}_{it}, \tau_i^2 \mathbb{I}_d), \\ &\sigma_{0i}^2 \sim \text{Inverse-Gamma}\left(a_{\sigma_0}, b_{\sigma_0}\right), && \tau_i^2 \sim \text{Gamma}\left(c_\tau, d_\tau\right), \end{aligned} \tag{4}$$

for $i = 1, ..., n\,;\, t = 1, ..., T-1$. Due to space constraints, we present the computation and theoretical results for such nodewise adaptive priors (4) in Section A.7 of the supplementary material.

In summary, the contributions of our paper can be summarized as follows:

1. Through the use of an SMF variational family, we proposed a CAVI algorithm, which offers an improvement over MF variational inference with minimal additional computational cost. Although our work was developed concurrently with Loyal and Chen (2023), which also adopts an SMF variational family, our approach is distinct in that it utilizes message passing rather than a variational Kalman smoother as in Loyal and Chen (2023);

2. A detailed theoretical analysis of the lower bound for the squared error loss associated with sufficiently smooth latent variables is presented, which is a first practice for dynamic latent space models to our best knowledge. In addition, we develop contraction rates of the posterior and its variational approximation of the proposed SMF procedure. To our best knowledge, such an analysis is the first result in the literature on Bayesian dynamic latent space models;

3. This technique for analyzing MF variational families for hierarchical prior distributions contributes to filling a gap in the recent literature regarding the analysis of variational inference for hierarchical prior distributions.

Notation. For a vector $\boldsymbol{x}$, we use $\|\boldsymbol{x}\|_2$, $\|\boldsymbol{x}\|_1$, $\|\boldsymbol{x}\|_\infty$ to represent its ℓ_2, ℓ_1 and ℓ_∞ norms and $\boldsymbol{x}'$ as its transpose. For a matrix $\boldsymbol{A}$, let $\|\boldsymbol{A}\|_F$ be its Frobenius norm. We use $\mathbb{I}$ and $\mathbf{1}$ to denote the identity matrix and vector with all ones. Suppose P and Q are probability measures on a common probability space with a dominating measure μ, and let $p = dP/d\mu, q = dQ/d\mu$. We use $D_{KL}\{p \,||\, q\} = \int p \log(p/q) d\mu$ to denote the KL divergence between the density p and q. In addition, we use $D_\alpha\{x \,||\, x_0\} = \log \int p_x^\alpha p_{x_0}^{1-\alpha} d\mu$ to denote the Rényi divergence of order α between the density p_x and p_{x_0}. Given sequences a_n and b_n, we denote $a_n = O(b_n)$ or $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \leq C b_n$ for all large enough n. Similarly, we define $a_n \gtrsim b_n$. In addition, let $a_n = o(b_n)$ to be $\lim_{n\to\infty} a_n/b_n = 0$. Let P_X denote a probability distribution with parameter X, and p_X denote the corresponding density function. Denote $\mathbf{E}_x$ as the expectation taken with respect to a variable x. Let $\mathcal{N}(\mu, \sigma)$ be the normal distribution with mean μ and variance σ while $N(x; \mu, \sigma)$ be the normal density function of value x with mean μ and variance σ. For any subset of B of Θ, we use $\Pi(B)$ to denote the probability of prior distribution taken on the set B.

2. Posterior Convergence of Dynamic Latent Space Models

Our main objective is to theoretically analyze the proposed SMF variational inference scheme. To this end, we first need to determine the appropriate parameter space to ensure that the variational posterior can converge at a near minimax rate. Since variational inference is an approximation of posterior inference, we will first establish that the α-fractional posterior (a variant of posterior) can achieve the minimax optimal rate for some specific parameter space. This will provide us with the necessary tools to prove the properties of the variational posterior. One of the sufficient conditions for optimal concentration of the α-variational posterior is that the α-posterior itself be well behaved (Yang et al., 2020), which in turn is guaranteed by Bhattacharya et al. (2019) through the optimal prior mass condition. In addition, the α-fractional posterior can indicate the optimal rate at which the upper-level parameters, such as scales, need to concentrate.

In this section, we first identify a suitable parameter space (7) for the unknown latent positions and obtain an information-theoretic lower bound to the rate of recovery (relative to a loss function defined subsequently) for said parameter space in Theorem 2. Such minimax lower-bound results for dynamic networks are scarce, and therefore this may be of independent interest. Next, under mild conditions on the evolution of the latent positions, we show in Theorem 3 that the rate of contraction of the fractional posterior matches the lower bound. We expand the posterior convergence result to include a prior on the transition variance.

2.1 Modeling Framework

We first state our assumptions on the data-generating process. Assume data is generated according to (1) with true latent position $\mathcal{X}^* = \{\boldsymbol{X}_t^*\}_{t=1}^T$. We assume that $\beta^* = 0$ for simplicity in theoretical analysis. Therefore, the network is assumed to be dense. As β in Equation (1) is unknown and approaches negative infinity for sparse networks, further research is needed to develop theoretical results for sparse networks in the Bayesian latent space model. We consider Gaussian or Bernoulli distributions for the observed links,

respectively,

$$\begin{aligned} &Y_{ijt} \overset{ind.}{\sim} \mathcal{N}(\boldsymbol{x}_{it}^{*'}\boldsymbol{x}_{jt}^{*}, \sigma^2), \quad 1 \le i \ne j \le n,\ t = 1, ..., T, \quad \text{or} \\ &Y_{ijt} \overset{ind.}{\sim} \text{Bernoulli}\left[1/\{1 + \exp(-\boldsymbol{x}_{it}^{*'}\boldsymbol{x}_{jt}^{*})\}\right], \quad 1 \le i \ne j \le n,\ t = 1, ..., T, \end{aligned} \tag{5}$$

where $\boldsymbol{x}_{it}^{*} \in \mathbb{R}^d$ is the ith row of $\boldsymbol{X}_t^{*}$ and designates the true latent coordinate of individual i at time t. The Gaussian likelihood can be considered a natural Bayesian alternative for estimating low-rank latent positions through singular value decomposition (SVD). It is worth noting that the optimization objective of SVD is associated with the best low-rank approximation in terms of the *Frobenius norm*, which implies that a Gaussian-likelihood type of Bayesian alternative can be used. On the other hand, the Bernoulli likelihood, which is a natural way to model binary responses, is widely used in the network and dynamic literature; see, for instance, Hoff et al. (2002); Sewell and Chen (2015); Ma et al. (2020); Zhang et al. (2020).

Fractional posterior: We adopt the expanded framework of a fractional posterior (Walker and Hjort, 2001), where the usual likelihood $P(\mathcal{Y} \mid \mathcal{X})$ is raised to a power $\alpha \in (0, 1)$ to form a pseudo-likelihood $P_\alpha(\mathcal{Y} \mid \mathcal{X}) := [P(\mathcal{Y} \mid \mathcal{X})]^\alpha$, which then leads to a fractional posterior $P_\alpha(\mathcal{X}, \tau, \sigma_0 \mid \mathcal{Y}) \propto P_\alpha(\mathcal{Y} \mid \mathcal{X})\, p(\mathcal{X} \mid \tau, \sigma_0) p(\tau) p(\sigma_0)$. Denote the ϵ ball for KL divergence neighborhood centered at $\mathcal{X}^*$ as

$$\begin{aligned} B_{n,T}(\mathcal{X}^*; \epsilon) := \Big\{ \mathcal{X} \in \Theta : &\int p_{\mathcal{X}^*} \log(\frac{p_{\mathcal{X}^*}}{p_{\mathcal{X}}}) d\mu \le n(n-1)T\epsilon^2, \\ &\int p_{\mathcal{X}^*} \log^2(\frac{p_{\mathcal{X}^*}}{p_{\mathcal{X}}}) d\mu \le n(n-1)T\epsilon^2 \Big\}, \end{aligned} \tag{6}$$

where μ is the Lebesgue measure and Θ is the parameter space of $\mathcal{X}$. Consider a subset B of the parameter space Θ, we use $\Pi_\alpha(B \mid \mathcal{Y}) = \int_B [P(\mathcal{Y} \mid \mathcal{X})]^\alpha P(X) d\mathcal{X} / \{\int_\Theta [P(\mathcal{Y} \mid \mathcal{X})]^\alpha P(X) d\mathcal{X}\}$ to denote the α-fractional posterior. Then our technique of analyzing fractional posterior is the following Lemma, adapted from Theorem 3.1 in Bhattacharya et al. (2019):

Lemma 1 (Contraction of fractional posterior distributions) *Fix $\alpha \in (0, 1)$. Assume $\epsilon_{n,T}$ satisfies $n^2 T \epsilon_{n,T}^2 \ge 2$ and*

$$\Pi\left(B_{n,T}\left(\mathcal{X}^*, \epsilon_{n,T}\right)\right) \ge e^{-n^2 T \epsilon_{n,T}^2}.$$

Then, for any $D \ge 2$ and $t > 0$,

$$\Pi_\alpha\left(\frac{1}{n(n-1)T} D_\alpha\left(\mathcal{X}, \mathcal{X}^*\right) \ge \frac{D + 3t}{1 - \alpha} \epsilon_{n,T}^2 \mid \mathcal{Y}\right) \le e^{-tn^2 T \epsilon_{n,T}^2}$$

holds with $P_{\mathcal{X}^*}$ probability at least $1 - 2/\left\{(D - 1 + t)^2 n^2 T \epsilon_{n,T}^2\right\}$.

In contrast to the theory of original posterior distributions, which requires additional conditions (for details, refer to Ghosal et al., 2000), the prior mass condition (6) is sufficient to ensure optimal concentration of fractional posterior. This is advantageous in the theoretical analysis of fractional posteriors because verifying the other conditions for complex

parameter spaces can be a challenging exercise. On the other hand, the fractional power α only appears as a multiple factor and will not affect the main rate with respect to n and T. In related literature, the α-variational posterior has been considered instead of the original one to facilitate theoretical analysis (see, for example, Linero and Yang, 2018; Martin and Tang, 2020; Jeong and Ghosal, 2021; Liu and Chen, 2022). They have also concluded that the concentration rates do not vary based on the choice of α. When $\alpha = 1/2$, the Hellinger divergence $h(p, q)$ is commonly used, and it is related to the $D_{1/2}(p, q) = -2\log(1-h^2(p, q))$, as discussed in Section 2.2 of Bhattacharya et al. (2019).

2.2 Lower Bounds to the Risk

We first examine the optimal lower bound under a suitable parameter space. To capture a smooth evolution of the latent coordinates over time, we assume the following parameter space for the latent position matrices:

$$\mathrm{PWD}(L) := \left\{ \mathcal{X} : \sum_{t=2}^{T} \sum_{i=1}^{n} \|\boldsymbol{x}_{it} - \boldsymbol{x}_{i(t-1)}\|_2 \leq L \right\}. \tag{7}$$

Here, PWD abbreviates point-wise dependence. The quantity L; which may depend on n or T; provides an aggregate quantification of the overall 'smoothness' in the evolution of the latent coordinates.

Given an estimator $\hat{\mathcal{X}}$ of $\mathcal{X}^*$, we consider the squared loss to formulate the minimax lower bound. Observe that the latent positions are only identifiable up to rotation, and thus the loss function above is formulated in terms of the Gram matrix corresponding to the latent position matrix, which is rotation invariant.

Theorem 2 (Minimax lower bound) *Suppose the data generating process follows Equation* (5)*. For* $\mathcal{X} \in \mathrm{PWD}(L)$*, with* $n - d + 1 \geq 16$*,* $n \geq 2d$*,* $T \geq 4$*, and* d *fixed, we have:*

$$\inf_{\hat{\mathcal{X}}} \sup_{\mathcal{X} \in \mathrm{PWD}(L)} \mathbf{E}_{\mathcal{X}} \left[\frac{1}{n(n-1)T} \sum_{t=1}^{T} \sum_{i \neq j=1}^{n} \left(\hat{\boldsymbol{x}}_{it}' \hat{\boldsymbol{x}}_{jt} - \boldsymbol{x}_{it}' \boldsymbol{x}_{jt} \right)^2 \right] \gtrsim \min \left\{ \frac{L^{\frac{2}{3}}}{n^{\frac{4}{3}} T^{\frac{2}{3}}}, \frac{1}{n} \right\} + \frac{1}{nT}.$$

While there is a sizable literature on minimax lower bounds for various static network models (Abbe and Sandon, 2015; Gao et al., 2015, 2016; Zhang and Zhou, 2016; Klopp et al., 2017), similar results for dynamic networks are scarce. To the best of our knowledge, only Pensky (2019) conducted such an analysis for dynamic stochastic block models, and there are no such results for latent space models. We prove the lower bound using a construction of a subset of low-rank latent states in Equation (A.1) in the appendix, which is adapted from the general construction of rank-one estimation of low-rank decomposed models (Vu and Lei, 2012; Birnbaum et al., 2013) to account for the network structure. We believe that such a construction can be used to analyze other latent space models for networks.

Theorem 2 characterizes the dependence of the lower bound on the number of time points T, the size of the network n, and the smoothness parameter L. We assume the latent dimension d to be a fixed constant in our calculations and refrain from making the dependence of the lower bound on d explicit. For fixed n, T, the term $L^{2/3} n^{-4/3} T^{-2/3}$ is an increasing function of L, implying that smoother transitions lead to better rates. However,

the rate cannot be faster than $1/(nT)$ even if L is arbitrarily small because under the extreme situation where all the latent positions $\boldsymbol{X}_1, ..., \boldsymbol{X}_T$ are the same, we still need to estimate a matrix of latent positions $\boldsymbol{X}_1$ with $O(n)$ parameters given $O(n^2T)$ observations. On the other hand, if L is large enough so that $T\sqrt{n}/L = o(1)$, the lower bound is $1/n$, which is equivalent to estimating each network separately ignoring the dependence. Finally, if n is fixed, the lower bound as a function of L and T reduces to $O(L^{2/3}T^{-2/3})$, which is the minimax rate for total variation denoising (Donoho and Johnstone, 1998; Mammen and van de Geer, 1997).

2.3 Convergence Rates of Fractional Posterior and Variational Risk

In this subsection, we show that under mild additional conditions, the minimax lower bound can be matched by the fractional posterior under the Gaussian random walk prior (2). First, we impose an identifiability condition in terms of a norm restriction:

$$\|\boldsymbol{x}_{it}^*\|_2 \leq C, \forall\, i = 1, ..., n, t = 1, ..., T, \text{ for some constant } C > 0. \tag{8}$$

Condition (8) requires that all the latent positions are norm-bounded by a constant, which is mild and reasonable considering the loss is in the inner product form. Under the above condition (8), all the probabilities induced by the inner product $p_{x_{it}^*, x_{jt}^*} := 1/\{1+\exp(-\boldsymbol{x}_{it}^{*'}\boldsymbol{x}_{jt}^*)\}$ are bounded away from 0 and 1 for the Bernoulli likelihood. Such an assumption is common for logistic models.

We additionally assume a homogeneity condition where we require that there exists a constant $C_0 > 0$, such that

$$\|\boldsymbol{x}_{it}^* - \boldsymbol{x}_{i(t-1)}^*\|_2 \leq C_0 L/(nT), \forall\, i = 1, ..., n, t = 1, ..., T. \tag{9}$$

If the true transitions satisfy (9), it is immediate they lie in the PWD class defined in (7). The homogeneity condition is compatible with random generating processes in the literature (Sewell and Chen, 2015)) such as a Gaussian random walk with bounded transition variance. Indeed, as long as $X_{ijt}^* - X_{ij(t-1)}^*$ for all i, j, t are sub-Gaussian random variables centered at zero and sub-Gaussian norm bounded by τ^*, using a concentration inequality for the maximal of sub-Gaussian random variables (Lemma 17), we have

$$P\left(\max_{i,j,t} |X_{ijt}^* - X_{ij(t-1)}^*| \geq \sqrt{2\tau^{*2}\{\log(nTd) + t\}}\right) \leq 2e^{-t}. \tag{10}$$

Therefore, with probability $1 - 2/(nTd)$, the homogeneity condition (9) holds when $\tau^* \leq C_0 L/(4nT\log(nTd))$. Similar conditions amounting to smooth transitions of the edge probabilities in a dynamic stochastic block model can also be found in Pensky (2019).

When considering binary likelihood and aiming for the convergence of the Frobenius norm of the difference between inner products, we also have a technical assumption for the prior: consider the event $B_p = \{\|\boldsymbol{x}_{it}\|_2 \leq C_4, \forall i \neq j = 1, ..., n, t = 1, ..., T\}$ for a constant $C_4 > \max \|\boldsymbol{x}_{it}^*\|_2$.

$$\text{We consider the prior restricted on event } B_p(\mathcal{X}),\ \tilde{\Pi} := \Pi(\cdot \cap B_p(\mathcal{X}))/\Pi(B_p(\mathcal{X})), \tag{11}$$

to replace the original prior such that $\tilde{\Pi}(B_p^c) = 0$. Without ambiguity, we still use $\Pi(\cdot)$ to denote $\tilde{\Pi}()$. However, the assumption is used for technical simplicity in the proof of the

Theorem, ensuring the connecting probabilities are controlled at a specific rate as in the literature (e.g., Ma et al., 2020; Zhang et al., 2022a,b), while not used in the algorithm. In particular, adopting such a restriction will only result in a negligible difference between using the original prior. A similar phenomenon is also reported in Remark 2 in Ma et al. (2020).

Under the above conditions, we have the following theorem:

Theorem 3 (Fractional posterior convergence with the fixed hyperparameters) *Suppose the true data generating process satisfies Equation* (5)*, $\mathcal{X}^* \in PWD(L)$ with $0 \leq L = o(Tn^2)$, and conditions* (8) *and* (9) *hold. Suppose d is a known fixed constant. Let $\epsilon_{n,T} = L^{1/3}/(T^{1/3}n^{2/3}) + \sqrt{\log(nT)/(nT)}$. Then, under the Gaussian random walk prior on $\mathcal{X}$ defined in Equation* (2) *and choosing σ_0 as a fixed constant and $\tau^2 = c_1\{\epsilon_{n,T}L/(nT) + \log^2(nT)/(nT^2)\}$ for some constants $c_1 > 0$; we have for $n, T \to \infty$,*

$$\Pi_\alpha\left(\frac{1}{n(n-1)T} D_\alpha\left(\mathcal{X}, \mathcal{X}^*\right) \geq M\epsilon_{n,T}^2 \mid \mathcal{Y}\right) \to 0.$$

In addition, if condition (11) *also holds, we also have*

$$\mathbf{E}\left[\Pi_\alpha\left\{\frac{1}{n(n-1)T}\sum_{t=1}^{T}\sum_{i\neq j=1}^{n}\left(\boldsymbol{x}_{it}'\boldsymbol{x}_{jt} - \boldsymbol{x}_{it}^{*\prime}\boldsymbol{x}_{jt}^*\right)^2 \geq M\epsilon_{n,T}^2 \mid \mathcal{Y}\right\}\right] \to 0, \tag{12}$$

with $P_{\mathcal{X}^*}$ probability converging to one, where $M > 0$ is a large enough constant.

Theorem 3 demonstrates that the minimax lower bound can be matched by the fractional posterior under specific choices of the hyperparameters σ and τ. In particular, the choice of τ ensues from an interplay between the smoothness of the Gaussian random walk prior and the truth. If τ is too small, the prior over-smoothes and fails to optimally capture the truth, while if τ is too large, then the prior under-smoothes, leading to overfitting. In particular, the smallest choice of τ^2 is at the rate of $\log^2(nT)/(nT^2)$, which corresponds to the smallest error rate $\sqrt{\log(nT)/(nT)}$. Moreover, Theorem 3 implies that when the dependence is weak (L is larger than $T\sqrt{n}$), applying Gaussian random walk priors with small transitions could damage the convergence rate of estimation accuracy. Besides, the rate implies that as long as the number of networks T is at least at the order of $L/\sqrt{n}$, the temporal dependence can be utilized to gain a rate no slower than the order of static network $\sqrt{1/n}$. The proof of Theorem 3 is based on transforming the Gaussian random walks into initial estimations together with Brownian motions initialed at zero and traditional techniques of calculating the shifted small ball probability for Brownian motions (e.g., Van der Vaart and Van Zanten, 2008).

Theorem 4 (Fractional posterior convergence with hierarchical priors) *Suppose the true data generating process satisfies Equation* (5)*, $\mathcal{X}^* \in PWD(L)$ with $0 \leq L = o(Tn^2)$, and conditions* (8) *and* (9) *hold. Suppose d is a known fixed constant. Let $\epsilon_{n,T} = L^{1/3}/(T^{1/3}n^{2/3}) + \sqrt{\log(nT)/(nT)}$. Then, under the Gaussian random walk prior on $\mathcal{X}$ defined in Equation* (2) *and adopting priors* (3) *for σ_0 and τ, we have for $n, T \to \infty$,*

$$\Pi_\alpha\left(\frac{1}{n(n-1)T} D_\alpha\left(\mathcal{X}, \mathcal{X}^*\right) \geq M\epsilon_{n,T}^2 \mid \mathcal{Y}\right) \to 0.$$

In addition, if condition (11) *also holds, we also have*

$$\mathbf{E}\left[\Pi_\alpha\left\{\frac{1}{n(n-1)T}\sum_{t=1}^{T}\sum_{i\neq j=1}^{n}\left(\boldsymbol{x}_{it}'\boldsymbol{x}_{jt}-\boldsymbol{x}_{it}^{*\prime}\boldsymbol{x}_{jt}^{*}\right)^2\geq M\epsilon_{n,T}^2\mid\mathcal{Y}\right\}\right]\to 0, \tag{13}$$

with $P_{\mathcal{X}^*}$ probability converging to one, where $M>0$ is a large enough constant.

Theorem 4, which is practically more relevant than Theorem 3, shows that the hierarchical prior on $\mathcal{X}$ specified by $\mathcal{X}\mid\sigma_0^2,\tau^2$ as in (2) and endowing the hyperparameters σ^2 and τ^2 with priors as in (3) leads to the same rate of contraction without knowledge of the smoothness parameter L. The Gamma prior on the transition variance τ^2 places sufficient mass around the 'optimal choice' in Theorem 3, which is a key ingredient in the proof of Theorem 4. We comment that the current proof technique does not work with an inverse-gamma prior on τ^2, with zero density at the origin.

3. Structured Mean-field in Latent Space Models

Variational approximations of fractional posteriors have also recently gained prominence (Bhattacharya et al., 2019; Alquier and Ridgway, 2020; Yang et al., 2020) — from a computational point of view, minor changes are needed while Bayes risk bounds for purely fractional powers ($\alpha<1$) require fewer conditions than the usual posterior ($\alpha=1$). Furthermore, as with the usual posterior, optimal convergence of the fractional posterior directly implies rate-optimal point estimators constructed from the fractional posterior. Variational inference approximates the posterior distribution $p(\mathcal{X},\beta,\tau,\sigma_0\mid\mathcal{Y})\propto P(\mathcal{Y}\mid\mathcal{X},\beta)\,p(\mathcal{X}\mid\tau,\sigma)p(\tau)p(\sigma_0)p(\beta)$ by its closest member in KL divergence from a pre-specified family of distributions Γ:

$$\begin{aligned}\hat{q}(\mathcal{X},\beta,\tau,\sigma_0)&=\underset{q(\mathcal{X},\beta,\tau,\sigma_0)\in\Gamma}{\operatorname{argmin}}\,D_{KL}\left\{q(\mathcal{X},\beta,\tau,\sigma_0)\,||\,p(\mathcal{X},\beta,\tau,\sigma_0\mid\mathcal{Y})\right\}\\&=\underset{q(\mathcal{X},\beta,\tau,\sigma_0)\in\Gamma}{\operatorname{argmin}}\,-\mathbf{E}_q\left\{\log\left(\frac{p(\mathcal{Y},\mathcal{X},\beta,\tau,\sigma_0)}{q(\mathcal{X},\beta,\tau,\sigma_0)}\right)\right\},\end{aligned} \tag{14}$$

where the term $\mathbf{E}_q\{\log(p(\mathcal{X},\beta,\tau,\sigma_0\mid\mathcal{Y})/q(\mathcal{X},\beta,\tau,\sigma_0))\}$ is called evidence-lower bound (ELBO).

3.1 The Structured Mean-field Family

For dynamic latent space models with fixed initial and transition scales σ_0 and τ, the mean-field (MF) variational family (Liu and Chen, 2022) assumes the form

$$q(\mathcal{X},\beta)=\left[\prod_{t=1}^{T}\prod_{i=1}^{n}q(\boldsymbol{x}_{it})\right]q(\beta). \tag{15}$$

The variational posterior under MF can be obtained through CAVI to maximize the ELBO (e.g., see Blei et al., 2017):

$$q^{(new)}(\beta)\propto\exp[\mathbf{E}_{-\beta}\{\log p(\mathcal{X},\beta,\mathcal{Y})\}];\quad q^{(new)}(\boldsymbol{x}_{it})\propto\exp[\mathbf{E}_{-\boldsymbol{x}_{it}}\{\log p(\mathcal{X},\beta,\mathcal{Y})\}], \tag{16}$$

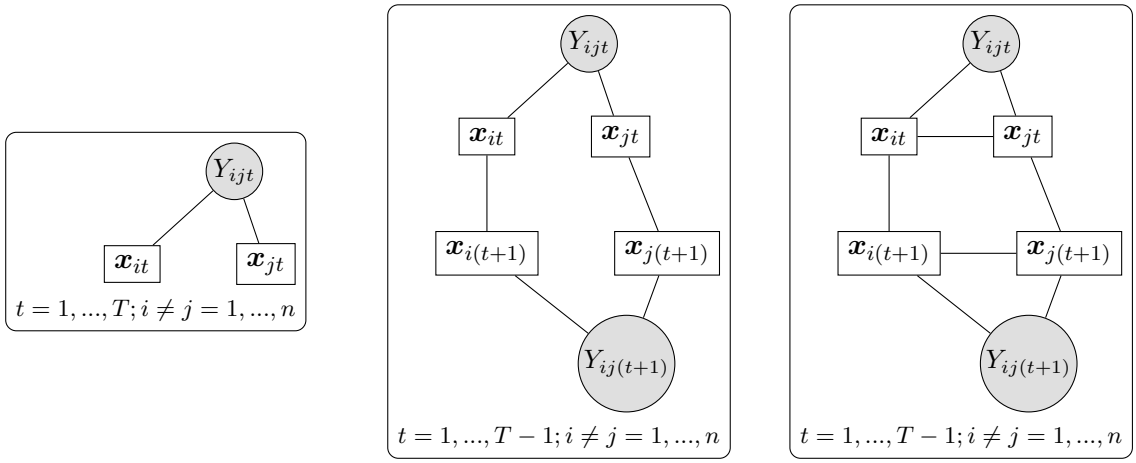

Figure 1: Graph representations for MF, SMF and exact posterior predictive distribution for latent space model given a fixed β, τ, σ_0. Conditional on $\mathcal{Y}$, the graph structure formed by latent positions are nT isolated nodes for MF, n separated chains with length T for SMF and a graph with many loops for posterior.

where $\mathbf{E}_{-\beta}$, $\mathbf{E}_{-\boldsymbol{x}_{it}}$ are the expectations taken with respect to the densities $q(\mathcal{X})$ and $\big[\prod_{t=1}^{T}\prod_{j\neq i} q(\boldsymbol{x}_{jt})\big]q(\beta)$, respectively. On the other hand, Sewell and Chen (2017) adopted a likelihood function where the dependence across time of $\mathcal{X}$ is captured by latent labels $\mathcal{Z}$. According to their Equation (22), they adopted a joint variational posterior for latent positions $q(\mathcal{X})$ with the assumption of independence among $\mathcal{X}$ in the variational family leading to independence of $\mathcal{X}$ across time $q(\mathcal{X}) = \prod_{t=1}^{T}\prod_{i=1}^{n} q(\boldsymbol{x}_{it})$. To see this more clearly, the algorithm in Section 2.2 of their supplementary material for the variational inference indicates to compute only all marginal distributions $q(\boldsymbol{x}_{it})$, rather than joint ones $q(\mathcal{X})$ in the variational posterior. Our proposed structured MF (SMF) variational family is instead given by

$$q(\mathcal{X}, \beta, \tau, \sigma_0) = \Big[\prod_{i=1}^{n} q_i(\boldsymbol{x}_{i\cdot})\Big]\, q(\beta)q(\tau)q(\sigma_0), \tag{17}$$

where $\boldsymbol{x}_{i\cdot} = [\boldsymbol{x}'_{i1}, \boldsymbol{x}'_{i2}, ..., \boldsymbol{x}'_{iT}]'$. Compared to MF, SMF does not enforce additional independence across time points $q_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)}) = q_{it}(\boldsymbol{x}_{it})q_{i(t+1)}(\boldsymbol{x}_{i(t+1)})$ for $i = 1, ..., n$, $t = 1, ..., T-1$. Figure 1 offers a visual comparison of the dependence structures among MF, SMF, and posterior predictives.

3.2 Computation for SMF

Utilizing the structure of the likelihood and prior, we have

$$\begin{aligned} &p_\alpha(\mathcal{Y}, \mathcal{X}, \beta, \tau, \sigma_0) \propto P_\alpha(\mathcal{Y} \mid \mathcal{X}, \beta)p(\mathcal{X} \mid \tau, \sigma_0)p(\tau)p(\sigma_0)p(\beta) \\ &= \prod_{t=1}^{T}\prod_{1\leq i\neq j\leq n} P_\alpha(Y_{ijt} \mid \boldsymbol{x}_{it}, \boldsymbol{x}_{jt}, \beta)\prod_{i=1}^{n}\left\{\prod_{t=1}^{T-1} p(\boldsymbol{x}_{i(t+1)} \mid \boldsymbol{x}_{it}, \tau)p(\boldsymbol{x}_{i1} \mid \sigma_0)\right\} \\ &\times p(\beta)p(\tau)p(\sigma_0), \end{aligned} \tag{18}$$

with $p(\boldsymbol{x}_{i(t+1)} \mid \boldsymbol{x}_{it}, \tau) \propto \exp(-\|\boldsymbol{x}_{i(t+1)} - \boldsymbol{x}_{it}\|_2^2/(2\tau^2))$ for $t = 1, ..., T-1$, where $\|\boldsymbol{x}\|_2$ represents its ℓ_2 norm of a vector $\boldsymbol{x}$. Based on the variational family (17), the CAVI updating of $q(\beta)$ $q(\tau)$,$q(\sigma_0)$ and $q_i(\boldsymbol{x}_{i\cdot})$, $i = 1, ..., n$ are performed in an alternating fashion. The update of β is standard and deferred to the supplemental material. We discuss the updating of the variance components in Section 3.5, and at present focus on the update of q_i. Specifically, suppose $q(\beta)$, $q(\tau)$,$q(\sigma_0)$ and $q_j(\boldsymbol{x}_{j\cdot})$, $j \neq i$ are fixed at their current values and the target is to update $q_i(\boldsymbol{x}_{i\cdot})$. The CAVI scheme gives

$$q_i(\boldsymbol{x}_{i\cdot}) \propto \exp\left[\mathbf{E}_{-\boldsymbol{x}_{i\cdot}}\left\{\sum_{t=1}^{T}\sum_{j\neq i, j=1}^{n}\{\log P_\alpha(Y_{ijt} \mid \boldsymbol{x}_{it}, \boldsymbol{x}_{jt}, \beta)\} + \sum_{t=1}^{T-1}\log p(\boldsymbol{x}_{i(t+1)} \mid \boldsymbol{x}_{it}, \tau) + \log p(\boldsymbol{x}_{i1} \mid \sigma_0)\right\}\right], \tag{19}$$

where $\mathbf{E}_{-\boldsymbol{x}_{it}}$ is the expectation taken with respect to the density $\big[\prod_{j\neq i} q_j(\boldsymbol{x}_{j\cdot})\big]q(\beta)q(\tau)q(\sigma_0)$.

Substituting the prior and likelihood (18) into Equation (19), it follows that $q_i(\boldsymbol{x}_{i\cdot})$ assumes the form:

$$q_i(\boldsymbol{x}_{i\cdot}) = q_{i1}(\boldsymbol{x}_{i1})\prod_{t=1}^{T-1} q(\boldsymbol{x}_{i(t+1)} \mid q(\boldsymbol{x}_{it})) = \prod_{t=1}^{T-1}\frac{q_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)})}{q_{it}(\boldsymbol{x}_{it})q_{i(t+1)}(\boldsymbol{x}_{i(t+1)})}\prod_{t=1}^{T} q_{it}(\boldsymbol{x}_{it}), \tag{20}$$

which implies that the graph of random variable $\boldsymbol{x}_{i\cdot}$ is structured by a chain from $\boldsymbol{x}_{i1}$ to $\boldsymbol{x}_{iT}$. It is important to notice that the structure (20) is not imposed by our variational family (17), rather a natural consequence of the Markov property of the prior and conditional independence of the likelihood in Equation (18). Given the above structure (20), computing the building blocks, i.e., the unary marginals $\{q_{it}\}$ and binary marginals $\{q_{it,i(t+1)}\}$, can be conducted in an efficient manner using message-passing (Pearl, 1982). To that end, we first define the following quantities:

$$\begin{aligned}
&\phi_{i1}(\boldsymbol{x}_{i1}) = \exp\{-\mu_{1/\tau^2}\|\boldsymbol{x}_{i1}\|_2^2/2 - \mu_{1/\sigma_0^2}\|\boldsymbol{x}_{i1}\|_2^2/2\}\prod_{j\neq i}\exp[\mathbf{E}_{q(\beta)q(\boldsymbol{x}_{j1})}\{\log P_\alpha(Y_{ij1} \mid \boldsymbol{x}_{i1}, \boldsymbol{x}_{j1}, \beta)\}],\\
&\phi_{it}(\boldsymbol{x}_{it}) = \exp\{-\mu_{1/\tau^2}\|\boldsymbol{x}_{it}\|_2^2/2\}\prod_{j\neq i}\exp[\mathbf{E}_{q(\beta)q(\boldsymbol{x}_{jt})}\{\log P_\alpha(Y_{ij1} \mid \boldsymbol{x}_{it}, \boldsymbol{x}_{jt}, \beta)\}],\ \forall t \in \{2, ..., T\}\\
&\psi_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)}) = \exp(\mu_{1/\tau^2}\boldsymbol{x}'_{i(t+1)}\boldsymbol{x}_{it}),\ \forall t \in \{1, ..., T-1\},
\end{aligned} \tag{21}$$

where $\mu_{1/\tau^2} = \mathbf{E}_{q(\tau)}(1/\tau^2)$ and $\mu_{1/\sigma_0^2} = \mathbf{E}_{q(\sigma_0)}(1/\sigma_0^2)$. For the ease of notation, we also denote $\psi_{i0,i1}(\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}) = 1$ and $\psi_{iT,i(T+1)}(\boldsymbol{x}_{iT}, \boldsymbol{x}_{i(T+1)}) = 1$.

Proposition 5 *The quantities appearing in the right-hand side of Equation* (20) *are given by,*

$$\begin{aligned}
q_{it}(\boldsymbol{x}_{it}) \ &\propto\ \phi_{it}(\boldsymbol{x}_{it})m_{i(t+1),it}(\boldsymbol{x}_{it})m_{i(t-1),it}(\boldsymbol{x}_{it}),\\
q_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)}) &\propto \phi_{it}(\boldsymbol{x}_{it})\phi_{i(t+1)}(\boldsymbol{x}_{i(t+1)})m_{i(t+2),i(t+1)}(\boldsymbol{x}_{i(t+1)})m_{i(t-1),it}(\boldsymbol{x}_{it}),
\end{aligned} \tag{22}$$

where

$$m_{i(t+1),it}(\boldsymbol{x}_{it}) \propto \int \phi_{i(t+1)}(\boldsymbol{x}_{i(t+1)})\psi_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)})m_{i(t+2),i(t+1)}(\boldsymbol{x}_{i(t+1)})d\boldsymbol{x}_{i(t+1)}$$

and

$$m_{it,i(t+1)}(\boldsymbol{x}_{i(t+1)}) \propto \int \phi_{it}(\boldsymbol{x}_{it})\psi_{it,i(t+1)}(\boldsymbol{x}_{it},\boldsymbol{x}_{i(t+1)})m_{i(t-1),it}(\boldsymbol{x}_{it})d\boldsymbol{x}_{it}$$

respectively are backward and forward messages for $t = 1, ..., T-1$.

In the message-passing literature, messages are computational items that can be reused from different marginalization queries, which are not necessary to be distributions (see Wainwright and Jordan, 2008 for more details). Proposition 5 provides the order of updatings to obtain $q_i(\boldsymbol{x}_{i\cdot})$: first, the initial backward/forward messages satisfy $m_{i(T+1),iT}(\boldsymbol{x}_{iT}) = m_{i0,i1}(\boldsymbol{x}_{i1}) = 1$. Then the other backward messages are obtained in the backward order from $m_{iT,i(T-1)}(\boldsymbol{x}_{i(T-1)})$ to $m_{i2,i1}(\boldsymbol{x}_{i1})$ and forward messages in the forward order from $m_{i1,i2}(\boldsymbol{x}_{i2})$ to $m_{i(T-1),iT}(\boldsymbol{x}_{iT})$. All messages are calculated based on the graph potentials in Equation (21), which can be computed analytically in conditionally conjugate Gaussian models illustrated in the next two subsections. Then updatings of all the unary and binary marginals are performed simultaneously according to Equation (22). Then the update of distribution $q(\boldsymbol{x}_{i\cdot})$ can also be obtained via property (20) thereafter.

The alternate MP updatings lead to an efficient block coordinate ascent algorithm where the dynamic structure of the same node is employed through MP within each block. When updating each node, the time complexity for MP is $O(T)$, hence the overall complexity per cycle is $O(nT)$. For linear state-space models, the established Kalman smoothing (Kalman, 1960) is often employed to obtain marginals of latent states efficiently. Our proposed algorithm is closely connected to Kalman smoothing. Specifically, we perform MP for a chain when updating each node, which is equivalent to Kalman smoothing for state-space models only up to updating rearrangements (Weiss and Pearl, 2010). Similar to the variational inference literature that uses Kalman smoothing in linear state-space models to replace MP (Barber and Chiappa, 2006), our proposed algorithm can also be rewritten as blockwisely implementing Kalman smoothing; see also Loyal and Chen (2023) for a parallel work in a hierarchical network model using the variational Kalman smoothing approach. We stick to the message-passing version of the proposed algorithm throughout the paper. Note that both the SMF and MF variational inference optimization problems are non-convex and can have many local minima. In order to make sure that the algorithms converge to a good optimum, multiple random starts can be used. Furthermore, SMF has an advantage over MF in that its optimization landscape may contain fewer local minima.

3.3 Gaussian Likelihood

We detail the steps of the SMF algorithm for a Gaussian likelihood:

$$P_\alpha(\mathcal{Y} \mid \mathcal{X}, \beta) = \prod_{t=1}^{T} \prod_{1\le i\ne j\le n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\alpha \frac{\{Y_{ijt} - \beta - \boldsymbol{x}_{it}'\boldsymbol{x}_{jt}\}^2}{2\sigma^2}\right].$$

where σ is assumed to be known. Suppose at current step, the variational distribution for node $\boldsymbol{x}_{it}$ follows the normal distribution $\mathcal{N}(\boldsymbol{\mu}_{it}, \boldsymbol{\Sigma}_{it})$, and the MF updating of $q(\beta)$ has been already performed so that $\mathbf{E}_{q(\beta)}(\beta) = \mu_\beta$ (provided in the supplementary material). Since $\phi_{it}(\boldsymbol{x}_{it})$ and $\psi_{it,i(t+1)}(\boldsymbol{x}_{it},\boldsymbol{x}_{i(t+1)})$ are proportional to Gaussian densities for

$\boldsymbol{x}_{it}$, the MP updating can be implemented in the framework of Gaussian belief propagation networks. Given node i, suppose $\phi_{it}(\boldsymbol{x}_{it})$ is proportional to $N(\boldsymbol{x}_{it}; \tilde{\boldsymbol{\mu}}_{it}, \tilde{\boldsymbol{\Sigma}}_{it})$, which is the density function of a $\mathcal{N}(\tilde{\boldsymbol{\mu}}_{it}, \tilde{\boldsymbol{\Sigma}}_{it})$ distribution evaluated at $\boldsymbol{x}_{it}$. Denoting $m_{it,i(t+1)}(\boldsymbol{x}_{i(t+1)}) \propto N(\boldsymbol{x}_{i(t+1)}; \boldsymbol{\mu}_{it\to i(t+1)}, \boldsymbol{\Sigma}_{it\to i(t+1)})$ and $m_{it,i(t-1)}(\boldsymbol{x}_{i(t-1)}) \propto N(\boldsymbol{x}_{i(t+1)}; \boldsymbol{\mu}_{it\to i(t-1)}, \boldsymbol{\Sigma}_{it\to i(t-1)})$, and based on calculations of Gaussian conjugate and marginalization using the Schur complement, we have the forward updating steps:

$$\boldsymbol{\mu}^{(new)}_{it\to i(t+1)} \leftarrow -\tau^2 \left[\tilde{\boldsymbol{\Sigma}}^{-1}_{it} \tilde{\boldsymbol{\mu}}_{it} + \boldsymbol{\Sigma}^{-1}_{i(t-1)\to it} \boldsymbol{\mu}_{i(t-1)\to it} + \alpha \sum_{j\neq i} (Y_{ijt} - \mu_\beta) \boldsymbol{\mu}_{jt} / \sigma^2 \right];$$

$$\boldsymbol{\Sigma}^{(new)}_{it,i(t+1)} \leftarrow -\tau^4 \left[\tilde{\boldsymbol{\Sigma}}^{-1}_{it} + \boldsymbol{\Sigma}^{-1}_{i(t-1)\to it} + \alpha \sum_{j\neq i} (\boldsymbol{\mu}_{jt}\boldsymbol{\mu}'_{jt} + \boldsymbol{\Sigma}_{jt}) / \sigma^2 \right]^{-1}.$$

Similarly, for the backward updating, we have

$$\boldsymbol{\mu}^{(new)}_{it\to i(t-1)} \leftarrow -\tau^2 \left[\tilde{\boldsymbol{\Sigma}}^{-1}_{it} \tilde{\boldsymbol{\mu}}_{it} + \boldsymbol{\Sigma}^{-1}_{i(t+1)\to it} \boldsymbol{\mu}_{i(t+1)\to it} + \alpha \sum_{j\neq i} (Y_{ijt} - \mu_\beta) \boldsymbol{\mu}_{jt} / \sigma^2 \right];$$

$$\boldsymbol{\Sigma}^{(new)}_{it,i(t-1)} \leftarrow -\tau^4 \left[\tilde{\boldsymbol{\Sigma}}^{-1}_{it} + \boldsymbol{\Sigma}^{-1}_{i(t+1)\to it} + \alpha \sum_{j\neq i} (\boldsymbol{\mu}_{jt}\boldsymbol{\mu}'_{jt} + \boldsymbol{\Sigma}_{jt}) / \sigma^2 \right]^{-1}.$$

3.4 Bernoulli Likelihood

Next, we consider a Bernoulli likelihood

$$P_\alpha(Y_{ijt} \mid \beta, \boldsymbol{x}_{it}, \boldsymbol{x}_{jt}) = \exp[\alpha Y_{ijt}(\beta + \boldsymbol{x}'_{it}\boldsymbol{x}_{jt}) - \alpha \log\{1 + \exp(\beta + \boldsymbol{x}'_{it}\boldsymbol{x}_{jt})\}],$$

where a larger value in $-\boldsymbol{x}'_{it}\boldsymbol{x}_{jt}$ results in a smaller probability that nodes i and j are connected at time t. We adopt the tangent transform approach proposed by Jaakkola and Jordan (2000) in the present context to obtain closed-form updates that are otherwise unavailable. The tangent-transform can be viewed as MF variational inference under Pólya–gamma data augmentation (Durante and Rigon, 2019). Statistical analysis of the tangent-transform for logistic regression was presented in Ghosh et al. (2022).

By introducing $\Xi = \{\xi_{ijt} : i, j = 1, ..., n, t = 1, ..., T\}$ with $A(\xi_{ijt}) = -\tanh(\xi_{ijt}/2)/(4\xi_{ijt})$ and $C(\xi_{ijt}) = \xi_{ijt}/2 - \log(1 + \exp(\xi_{ijt})) + \xi_{ijt}\tanh(\xi_{ijt}/2)/(4\xi_{ijt})$ for any ξ_{ijt}, we have the following lower bound on $P_\alpha(Y_{ijt} \mid \boldsymbol{x}_{it}, \boldsymbol{x}_{jt}, \beta)$:

$$\underline{P}_\alpha(Y_{ijt} \mid \beta, \boldsymbol{x}_{it}, \boldsymbol{x}_{jt}; \xi_{ijt}) = \exp\left[\alpha A(\xi_{ijt})(\beta + \boldsymbol{x}'_{it}\boldsymbol{x}_{jt})^2 + \alpha \left(Y_{ijt} - \frac{1}{2} \right) (\beta + \boldsymbol{x}'_{it}\boldsymbol{x}_{jt}) + \alpha C(\xi_{ijt}) \right].$$

By replacing $P_\alpha(Y_{ijt} \mid \boldsymbol{x}_{it}, \boldsymbol{x}_{jt}, \beta)$ with its lower bound $\underline{P}_\alpha(Y_{ijt} \mid \boldsymbol{x}_{it}, \boldsymbol{x}_{jt}; \xi_{ijt}, \beta)$, we can update the posterior distribution of $\mathcal{X}$ in the Gaussian conjugate framework given the rest densities. After updating all the blocks, ξ_{ijt} is optimized based on EM algorithm and

the property of $A(\xi_{ijt})$ according to Jaakkola and Jordan (2000): $\xi_{ijt}^{(new)2} = \mathbf{E}_{q(\beta,\mathcal{X})}\{(\beta + \boldsymbol{x}_{it}'\boldsymbol{x}_{jt})^2\}$.

In summary, for Gaussian or Bernoulli likelihood, the SMF framework allows all updatings in the Gaussian conjugate paradigm by only assuming independence between different nodes in the variational family.

3.5 Updatings of Scales

The updating of scales can also be performed in closed form. Note that with the Gamma(c_τ, d_τ) prior for τ, by the CAVI algorithm, we have

$$q(\tau^2) \propto \exp\left[\mathbf{E}_{q(\mathcal{X})}\left\{-\sum_{t=2}^{T}\frac{\|\boldsymbol{X}_t - \boldsymbol{X}_{t-1}\|_F^2}{2\tau^2}\right\} - \frac{n(T-1)d + c_\tau - 1}{2}\log(\tau^2) - d_\tau\tau^2\right]. \quad (23)$$

Equation (23) implies that the new update of τ^2 under CAVI follows a Generalized inverse Gaussian distribution (Jorgensen, 2012) with parameter $a = 2d_\tau, b = \mathbf{E}_{q(\mathcal{X})}\{\sum_{t=2}^{T}\|\boldsymbol{X}_t - \boldsymbol{X}_{t-1}\|_F^2/2\}, p = 1/2 - n(T-1)d/2 - c_\tau/2$, where $\|\cdot\|_F$ is the Frobenius norm. Then the moment required in updating $\mathcal{X}$ in Equation (21) can be obtained: $\mathbf{E}_{q(\tau)}(1/\tau^2) = \sqrt{a}K_{p+1}(\sqrt{ab})/\left\{\sqrt{b}K_p(\sqrt{ab})\right\} - 2p/b$, where $K_p(\cdot)$ is the modified Bessel function of the second kind. When p is large, overflow could happen in directly calculating the value of $K_p(\cdot)$. To address this issue, expansions of $K_p(\cdot)$ can be performed in the logarithm scale, which is implemented in R package *Bessel* (Maechler, 2019).

For the initial variance σ_0 with prior (3), the inverse-Gamma conjugate updating can be performed:

$$q(\sigma_0^2) \propto \exp\left[\mathbf{E}_{q(\mathcal{X})}\left(-\frac{\|\boldsymbol{X}_1\|_F^2}{2\sigma_0^2}\right) - \left(\frac{nd}{2} + a_{\sigma_0} + 1\right)\log(\sigma_0^2) - \frac{b_{\sigma_0}}{\sigma_0^2}\right]. \quad (24)$$

Hence we have $\sigma_0^{(new)2} \sim$ Inverse-Gamma$((nd + a_{\sigma_0})/2, \{\mathbf{E}_{q(\mathcal{X})}(\|\boldsymbol{X}_1\|_F^2) + 2b_{\sigma_0}\}/2)$, which implies $\mu_{1/\sigma_0^2} = \mathbf{E}_{q(\sigma_0)}(1/\sigma_0^2) = (nd + a_{\sigma_0})/\{\mathbf{E}_{q(\mathcal{X})}(\|\boldsymbol{X}_1\|_F^2) + 2b_{\sigma_0}\}$.

The choice of the priors (3) of the scales leads to both the closed-form updating algorithm in CAVI and the optimal convergence rate detailed in the next section. Finally, it is important to notice that the above computational framework can be safely extended to nodewise adaptive priors defined in Equation (4), whose details are in Section A.7 in the supplementary material.

3.6 Theoretical Results for SMF

To show the theoretical result of the global optimizer of the proposed SMF algorithm. First, we need the following Lemma, which is adapted from Lemma 3.3 from Yang et al. (2020) to prove the convergence of the α-fractional variational posterior:

Lemma 6 (Variational risk bound of the α-fractional variational posterior) *With $P_{\mathcal{X}^*}$ probability at least $1-\xi$ that for any probability measure $q_{\mathcal{X}} \ll p_{\mathcal{X}}$, we have*

$$\int \{D_\alpha(X\|\mathcal{X}^*)\}\,\hat{q}_{\mathcal{X}}(\mathcal{X})d\mathcal{X}$$
$$= \frac{\alpha}{n(n-1)T(1-\alpha)}\left[-\int_\Theta \log\frac{p(Y \mid \mathcal{X})}{p(Y \mid \mathcal{X}^*)}q_{\mathcal{X}}(\mathcal{X})d\mathcal{X} + \frac{D(q_{\mathcal{X}}\|p_{\mathcal{X}})}{\alpha} + \frac{\log(1/\zeta)}{\alpha}\right],$$

where $\hat{q}_{\mathcal{X}}(\mathcal{X})$ is the global minimizer of the KL divergence under the α-fractional framework.

Finally, we show in Theorems 7 and 8 below that the Bayes risk bound from Theorem 3 and 4 is retained under the optimal SMF solution $\hat{q}$ by using Lemma 6. As an important upshot, the point estimate obtained from the variational solution retains the same convergence rate as the fractional posterior.

Theorem 7 (Variational risk bound for marginal VB families) *Suppose the true data generating process satisfies Equation* (5)*, $\mathcal{X}^* \in PWD(L)$ with $0 \leq L = o(Tn^2)$ and conditions* (8) *and* (9) *hold. Suppose d is a known fixed constant. Let $\epsilon_{n,T} = L^{1/3}/(T^{1/3}n^{2/3}) + \sqrt{\log(nT)/nT}$. Then if we apply the priors defined in Equation* (2)*, and either the following* (*a*) *or* (*b*) *holds:*

(a). choosing σ_0 as a fixed constant and $\tau^2 = c_1\{\epsilon_{n,T}L/(nT) + \log^2(nT)/(nT^2)\}$ for some constants $c_1 > 0$ and obtaining the optimal variational distribution $\hat{q}(\mathcal{X})$ under the SMF family $q(\mathcal{X}) = \prod_{i=1}^n q_i(\boldsymbol{x}_{i\cdot})$;

(b). adopting priors (3) *for σ_0 and τ and obtaining the optimal variational distribution $\hat{q}(\mathcal{X})$ under the SMF family $q(\mathcal{X}) = \prod_{i=1}^n q_i(\boldsymbol{x}_{i\cdot})$;*

we have with $P_{\mathcal{X}^*}$ probability tending to one as $n, T \to \infty$,

$$\int \frac{1}{n(n-1)T} D_\alpha\left(\mathcal{X}, \mathcal{X}^*\right) \hat{q}(\mathcal{X}) d\mathcal{X} \lesssim \epsilon_{n,T}^2.$$

In addition, if condition (11) *also holds, we also have*

$$\mathbf{E}_{\hat{q}(\mathcal{X})}\left[\frac{1}{n(n-1)T}\sum_{t=1}^{T}\sum_{i\neq j=1}^{n}\left(\boldsymbol{x}_{it}'\boldsymbol{x}_{jt} - \boldsymbol{x}_{it}^{*\prime}\boldsymbol{x}_{jt}^*\right)^2\right] \lesssim \epsilon_{n,T}^2. \tag{25}$$

Theorem 7 (a) and (b) correspond to the strategies adopted in Liu and Chen (2022) and Bai et al. (2020) respectively. Theorem 7 (a) requires the tuning of hyperparameters, which loses the adaptive property of posterior under the adopted hierarchical prior as in Theorem 4. In Theorem 7 (b), the variational inference is performed within a marginal family, resulting in a richer family than additional independence between $\mathcal{X}$ and τ, σ_0 as in Equation (17). The optimization with respect to the marginal VI family, however, does not have an analytical expression and will therefore require Monte Carlo approximations (see discussions in Appendix C in Bai et al., 2020), which is not inconvenient as the algorithm proposed towards VI family (17). Nevertheless, our following Theorem 8 shows that the gap between computation and theory can be bridged.

Theorem 8 (Variational risk bound for SMF) *Suppose the true data generating process satisfies Equation* (5)*, $\mathcal{X}^* \in PWD(L)$ with $0 \leq L = o(Tn^2)$ and conditions* (8) *and* (9) *hold. Suppose d is a known fixed constant. Let $\epsilon_{n,T} = L^{1/3}/(T^{1/3}n^{2/3}) + \sqrt{\log(nT)/nT}$. Then if we apply the priors defined in Equation* (2)*, and adopt priors* (3) *for σ_0 and τ and obtaining the optimal variational distribution $\hat{q}(\mathcal{X})$ under SMF family* (17)*, we have with $P_{\mathcal{X}^*}$ probability tending to one as $n, T \to \infty$,*

$$\int \frac{1}{n(n-1)T} D_\alpha\left(\mathcal{X}, \mathcal{X}^*\right) \hat{q}(\mathcal{X}) d\mathcal{X} \lesssim \epsilon_{n,T}^2.$$

In addition, if condition (11) *also holds, we also have*

$$\mathbf{E}_{\hat{q}(\mathcal{X})}\left[\frac{1}{n(n-1)T}\sum_{t=1}^{T}\sum_{i\neq j=1}^{n}\left(\boldsymbol{x}_{it}'\boldsymbol{x}_{jt}-\boldsymbol{x}_{it}^{*\prime}\boldsymbol{x}_{jt}^{*}\right)^2\right]\lesssim\epsilon_{n,T}^2. \tag{26}$$

Theorem 8 is the main theorem in this paper that corresponds to the proposed algorithm. It implies that the independence in the SMF family between $\mathcal{X}$ and τ, σ_0 does not bring any damage to the convergence rate of the optimal variational estimator. Therefore, the proposed VI algorithm enjoys the same adaptive property as the posterior under the adopted hierarchical prior without any loss. The proof strategy of Theorem 8 has a key distinction with Theorem 7: the mismatch between the hierarchical prior $p(\mathcal{X}, \tau, \sigma_0) = p(\mathcal{X} \mid \tau, \sigma_0)p(\tau)p(\sigma_0)$ and independent variational family $q(\mathcal{X}, \tau, \sigma_0) = q(\mathcal{X})q(\tau)q(\sigma_0)$ adds some complexity in the analysis. We construct a candidate in the variational family, which leads to the optimal rate. Specifically, for any chosen τ^*, σ_0^* that satisfy the condition in Theorem 7 (a), construct $\bar{q}(\mathcal{X}, \tau, \sigma_0)$ by restricting $p(\boldsymbol{x}_{i(t+1)} \mid \boldsymbol{x}_{it}, \tau^*)$ to a neighborhood of $\boldsymbol{x}_{i(t+1)}^*$ for all i and $t > 1$ and restricting $p(x_{i1} \mid \sigma_0^*)$ to a neighborhood of $\boldsymbol{x}_{i1}^*$ for all i. Also restrict $p(\tau)$, $p(\sigma_0)$ to a neighborhood of τ^*, σ_0^* (see Equation (A.14) in the appendix). Observe that such a construction lies within the proposed SMF family (17). By an appropriate choice of the size of the neighborhood, we can achieve

$$D_{KL}(\bar{q}(\mathcal{X}, \tau, \sigma_0) \,||\, p(\mathcal{X}, \tau, \sigma_0)) \lesssim Tn(n-1)\epsilon_{n,T}^2.$$

In addition, by using the decomposition

$$\begin{aligned} D_{KL}(\bar{q}(\mathcal{X}, \tau, \sigma_0) \,||\, p(\mathcal{X}, \tau, \sigma_0)) = D_{KL}(\bar{q}(\tau)||p(\tau)) + D_{KL}(\bar{q}(\sigma_0)||p(\sigma_0)) \\ + \int \bar{q}(\tau)q(\bar{\sigma}_0)\int \bar{q}(\mathcal{X})\log\frac{\bar{q}(\mathcal{X})}{p(\mathcal{X}|\tau,\sigma_0)}d\mathcal{X}d\tau d\sigma_0, \end{aligned}$$

we are able to show that the selected candidate achieves optimal bounds for all three terms on the right-hand side. We conjecture that this strategy is fairly general and can be applied to mean-field inference for other Bayesian hierarchical models as well.

4. Simulations and Real Data Analysis

In this section, we will first provide simulation examples to illustrate our results. Then, we will present two real data examples: the Enron Email data set and the McFarland Classroom data set.

4.1 Simulation Experiments

We perform replicated simulation studies to compare SMF, MF and MCMC. Throughout all simulation and real data analyses, we fix the fractional power $\alpha = 0.95$. We also fix the hyperparameters $a_{\sigma_0} = 1/2, b_{\sigma_0} = 1/2$ and $c_\tau = 1, d_\tau = 1/2$ whenever the prior (3) is used. Simulation results for Gaussian likelihood can be found in Section A.9 in the appendix.

Binary Networks: 25 replicated data sets are generated from (1) with $Y_{ijt} \sim$ Bernoulli $[1/\{1 + \exp(-2 + \boldsymbol{x}_{it}'\boldsymbol{x}_{jt})\}]$ for $i \neq j = 1, ..., n$ and $t = 1, ..., T$ with $d = 2$. The latent positions are initialized as $\boldsymbol{x}_{i1} \sim 0.5\mathcal{N}((1, 0)', 0.5^2\mathbb{I}) + 0.5\mathcal{N}((-1, 0)', 0.5^2\mathbb{I})$ with subsequent

draws from $\boldsymbol{x}_{it} = \boldsymbol{x}_{i(t-1)} + \boldsymbol{\epsilon}_{t-1}$, where given any coordinate j for a fixed node i, we have $[\epsilon_{ij1}, ..., \epsilon_{ijT}]' \sim \mathcal{N}(\mathbf{0}, \tau^2((1-\rho)\mathbb{I} + \rho\mathbf{1}\mathbf{1}'))$. The transition sd τ controls the magnitude of transition, and the auto-correlation ρ controls the positive dependence. As a measure of discrepancy between the true and estimated probabilities, we use the sample Pearson correlation coefficient (PCC, which is also used in other literature, e.g, Sewell and Chen, 2017): $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}/\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})}$ for two lists of probabilities $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$. The number of iterations until convergence is reported to investigate the computational efficiency. The stopping criterion is taken to be the difference between training AUCs (area under the curve) in two consecutive cycles not exceeding 0.01. To implement SMF and MF, we assume the initial variance to be $\sigma_0 = 0.5$. The prior for parameter β is set to $\mathcal{N}(0, 10)$.

We compared standard MCMC and SMF in terms of estimation accuracy and computation time for binary networks. We ran MCMC with 100, 200 and 5000 iterations using a Gibbs Sampler algorithm, where each coefficient was sampled from its full conditional distribution. For the MCMC chain, we discarded the first half of iterations as burn-in and used the sample means from the last half of iterations to calculate the estimator. We then compared this accuracy with SMF using the PCC with the true probabilities, as well as considering computation time. We set the transition smoothness $\tau = 0.01, 0.05, 0.1$, sample size $n = 10, 20, 50$, time point $T = 100$, and correlation $\rho = 0.5$. The simulations were repeated 25 times for each setting. The results are presented in boxplot comparisons shown in Figure 2 and Figure 3, which illustrate several noteworthy findings: First, the stronger the dependence across time, the better the performance of SMF in terms of higher PCC accuracy when τ decreases from 0.1 to 0.01. This is because the computation of SMF incorporates the dependence across time, resulting in improved performance. However, for MCMC, the weaker the dependence across time, the better the performance in terms of higher PCC accuracy when τ increases from 0.01 to 0.1. This is because the mixing of the Markov Chain is affected negatively by the dependence across time. Similarly, for MF, the weaker the dependence across time, the better the performance in terms of higher PCC accuracy when τ increases from 0.01 to 0.1, as weaker dependence will better fit the independence structure of MF. Increasing τ reduces the gap in estimation accuracy for MCMC among iterations 100, 200 and 5000, indicating faster mixing of the Markov Chains. Overall, SMF requires less computation time than MCMC and MF under the given settings while achieving almost the best estimation accuracy, which is similar to MCMC with 5000 iterations. This indicates that when the dependence across time is strong, SMF significantly improves computation efficiency.

In Table 1, we report the median of PCC from 25 simulation experiments for binary networks with $n = 100$ and $T = 100$. The comparison is between the true and estimated probabilities for SMF with $\sigma_0 = 0.5$ and a known τ, versus adaptive SMF using (3). It's interesting to note that learning the initial and transition variances adaptively using the prior (3) doesn't lead to any loss of accuracy compared to when these parameters are known as a priori.

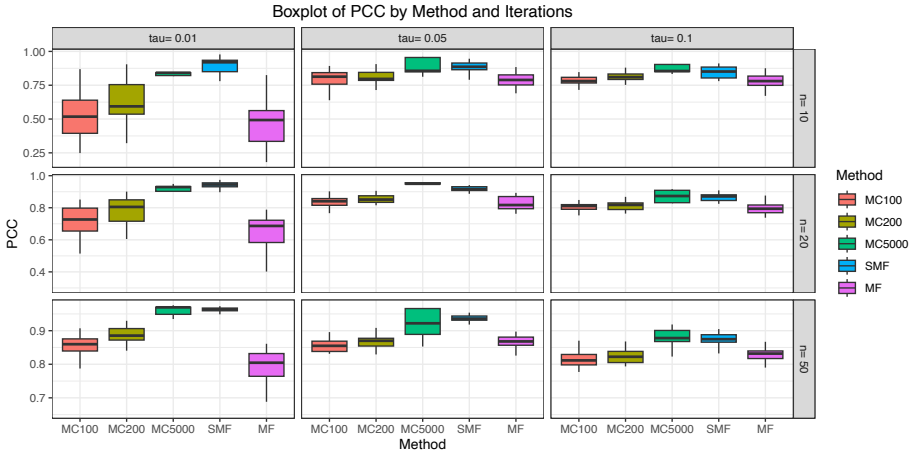

Figure 2: Boxplots comparing the estimation accuracy of Pearson Correlation Coefficient (PCC) between the estimated and true connected probabilities for SMF, MF, and various numbers of MCMC iterations. A higher PCC indicates better estimation performance for the corresponding method. MC100, MC200 and MC 5000 represent posterior means obtained after 100, 200 and 5000 iterations of Gibbs samplers, respectively, with the first half of iterations discarded as burn-in. Among all cases, SMF achieves a similar level of estimation accuracy with MCMC with 5000 iterations.

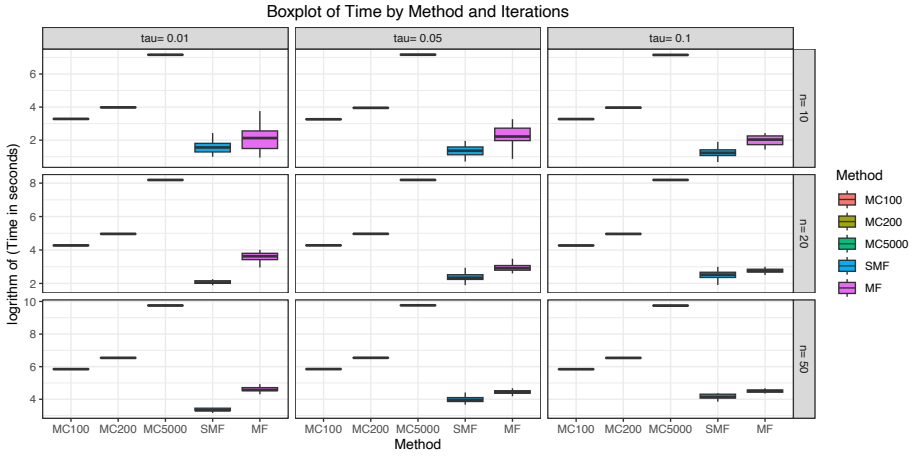

Figure 3: Boxplots comparing the computation time for SMF, MF, and different numbers of MCMC iterations, as simulated in Figure 2. MC100, MC200 and MC 5000 represent 100, 200 and 5000 iterations of the Gibbs samplers, respectively. Remarkably, SMF exhibits the shortest computation time while almost achieving the highest level of estimation accuracy in Figure 2 across all cases.

τ	0.01			0.1		
ρ	0	0.4	0.8	0	0.4	0.8
Adaptive SMF	0.885	0.888	0.892	0.880	0.910	0.914
SMF	0.887	0.881	0.898	0.897	0.912	0.904
τ	0.2			0.3		
ρ	0	0.4	0.8	0	0.4	0.8
Adaptive SMF	0.907	0.915	0.918	0.908	0.919	0.923
SMF	0.895	0.918	0.915	0.904	0.919	0.931

Table 1: Performance comparisons for binary networks between adaptive SMF and SMF with known initial and transition variances. The measure compared are the medians of Pearson correlation coefficient (PCC) between true and estimated probabilities of the repeated simulations.

4.2 Enron Email

Using the Enron email data set (Klimt and Yang, 2004), we compare our model with the latent space model with the same likelihood but with an inverse Gamma prior on the transition variance. Enron data consists of emails collected from 2359 employees of the Enron company. From all the emails, we examine a subset consisting of $n = 184$ employees communicating among $T = 44$ months from Nov. 1998 to June 2002 recorded in the `R` package `networkDynamic` (Butts et al., 2020). The networks depict the email communication status of employees over that period. The edges in the network are ones if one of the corresponding two employees sent at least one email to the other during that month. According to the data set, all networks are sparse and many edges remain unchanged over time. The aim of this study is to determine whether shrinkage on transitions induced by Gamma prior on transition variance can be beneficial for sparse dynamic networks. With the dynamic networks, we consider all the edges to be missed with probability $p = 0.01, 0.02, ..., 0.1$ independently, train the two latent space models without the missed data, and then make predictions based on the missed data. We use two criteria for comparison: the testing AUC score and the ratio of true positive detection over all missed edges, which is defined as the ratio of predictive probability greater than 0.5 when the true edge value is 1 over all missed edges. Since all networks are extremely sparse and negative predictions are trivial, the second criterion above is meaningful. The same SMF variational inference method is used in both latent space models. In both of the latent space models, we assign a latent dimension of 5 (more results about $d = 2, 3, 4$ are provided in Section A.9 in the appendix), the same initialization and stopping criteria. The variational mean of the latent positions is used to estimate the latent positions.

Figure 4 illustrates a performance comparison between the two approaches. The Gamma prior leads to a better fit based on the AUC comparison (left subfigure) and improves the detection of missed links (right subfigure). A Gamma prior shrinks the transitions more compared to an inverse-Gamma prior so that if two employees communicate at time t, the predictive probability for them to communicate at time $t + 1$ is high.

Figure 4: Comparisons of latent space models between Gamma or Inverse Gamma priors on the Enron email testing data.

4.3 McFarland Classroom

McFarland's streaming classroom data set provides interactions of conversation turns from streaming observations of a class observed by Daniel McFarland in 1996 (McFarland, 2001). The data set is available in the `R` package `networkDynamic` (Butts et al., 2020). The class comprised of 2 instructors and 18 students. Of the 2 instructors, one is the main instructor who lectured most of the time, while the other is an assistant. During the class, the instructors began by providing instructions to all students. Then, the students were divided into groups and assigned collaborative group work. The two instructors oversaw the activities across the groups to assist the students. Here, we aim to compare MF and SMF via prediction accuracy and visualize the dynamic evolution of the latent positions.

We divide the entire class time into 8 equispaced time points. The edges of each of the 8 networks represent whether the two nodes interacted related to the study task during the entire time period. We chose $d = 2$ for visualization purposes. A $\mathcal{N}(0, 10)$ prior is placed on the intercept and the prior (3) is adopted for both SMF and MF. First, we compare SMF and MF in terms of prediction accuracy. For $t = 3, 4, ..., 8$, the first $t-1$ networks are used as the training data, while the t-th network is used as the test data. The estimated latent positions at time point $t - 1$ are used to predict the probabilities of edges between any two nodes at time t. Then the test AUC scores are obtained from the above-estimated probabilities vs. the true binary responses at time point t. We repeat the process for 25 times with different initial values. The boxplots of the test AUC scores for MF and SMF are shown in Figure 5. From the figure, we can see that except for time point $t = 5$, where the network structure changed significantly and the dependence from previous time points may not be meaningful (see Figure 8 for the change of the connections), SMF consistently performs better than MF, which again testified to the ability for SMF to capture the dependence across time better.

Next, we implemented SMF with the networks at all 8 time points under the same hyperparameter specification to visualize the dynamic evolution of the latent positions. Since the latent positions estimated directly from the algorithm are not identifiable, Procrustes rotation is performed (Hoff et al., 2002) where the latent positions of time $t = 2, .., T$ are

projected to the locations that are most close to its previous locations ($t = 1, ..., T-1$) through Procrustes rotation. Observe that the inner product is invariant to this transformation. Figure 6 shows the dynamic evolution of the variational mean of the latent positions for both students and instructors (an animated version of Figure 6 is provided in the supplementary material). At time point 1 (i.e., at the beginning of the class) the students indexed by $\{1, ..., 20\}\backslash\{7, 14\}$ are approximately grouped into the following clusters $(6, 11, 15)$, $(3, 8, 13)$, $(10, 12, 4, 5)$, $(1, 18, 9)$, $(2, 19)$ and $(20, 17, 16)$. The locations of the students remained the same until time point 4. From time point 4 to 5, the inner-group distances between $(1, 9, 18)$ and $(4, 5, 10, 12)$ became smaller, which reflected the real scenario that the students were assigned into groups. Then the group structure of the students remained similar for the remainder of the class. Overall, the evolution reflected the collaborative behavior between certain groups as they performed specific tasks during the class. As a point of comparison, we also obtained dynamic visualization of the networks via MF (Figure 7) and the popular `ndtv` package (ndtv: Network Dynamic Temporal Visualization, Bender-deMoll and Morris, 2021). Although `ndtv` package is known for its dynamic networks visualizations through animations, static snapshots of the visualizations can also be created using `filmstrip` function (Figure 8). First, unlike Figure 6, the latent positions estimated via MF in Figure 7 did not have a smooth temporal evolution, as the MF assumed independence across the time points. In addition, compared to our visualization in Figure 6, results from the `ndtv` package in Figure 8 lacked a clear pattern of the network evolution. For example, the students indexed $\{1, 18, 9\}$ stayed close to each other at time points $t = 5, 6, 7$ in Figure 6, while in Figure 8, 18 is far away from $(1, 9)$ at time $t = 6$, while being connected to 1 at the neighboring time points $t = 5, 7$. A similar phenomenon can be seen for student indexed 5 at time $t = 5$, where in Figure 6 it is close to $(4, 10, 12)$ while in Figure 8 it is not. The ability of our methodology to borrow information across time is specifically due to the Markovian structure (2) imposed on the evolution of the latent positions endowed with the Gamma prior (3) on the transition variance, allowing sufficient probability near the origin. Thus our methodology revealed a more realistic pattern in the evolution in Figure 6 compared to MF and `ndtv` as most of the detected changes remained concentrated in time $t = 4, 5$ for the students (when the students formed groups) and $5, 6, 7$ for the instructors (after the instructors began assisting the students).

5. Discussion

There are a number of potential extensions of the proposed methodology and theory in this article. Properties of the Gaussian random walk prior is crucially exploited in our theory to obtain the optimal variational risk. It would be interesting to explore similar theoretical optimality results for Gaussian Process priors (e.g., Durante et al., 2017b). Moreover, the theoretical analysis of the lower bound can be extended to the case that the true latent positions evolve smoothly over time, like in Pensky (2019).

From a methodological point, it is of interest to explore how to perform community detection after estimating the latent positions. As the latent positions are characterized as vectors in Euclidean space, it is natural to consider some distance-based approaches like K-means for clustering. Adapting to the dimension d of the embedding space is also a challenging problem. Finally, it is also interesting to explore dynamic latent space models

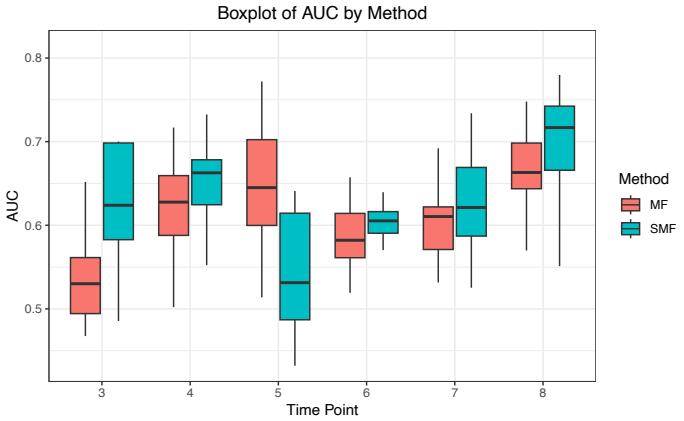

Figure 5: Boxplot comparing AUC Predictions between SMF and MF for Time Points $t = 3$ to 8, using different algorithm initializations based on networks from previous time points. SMF outperforms MF consistently, except at time point 5, where significant structural changes in the network may hinder the benefit of temporal dependencies across time points.

with other complex data to fit real-world scenarios, such as continuous-time networks (Loyal, 2024) and dynamic networks with sudden structural changes (Zhao et al., 2022).

Acknowledgments

We are grateful to the action editor and four reviewers for their valuable suggestions and kind help, which significantly improved the quality of the paper. The research is supported by the National Science Foundation CCF-1934904. Drs. Pati and Bhattacharya acknowledge support from NSF DMS 2210689.

Reproducible implementations and experiments are publicly available at `https://github.com/pengzhaostat/SMF-structured-variational-inference`.

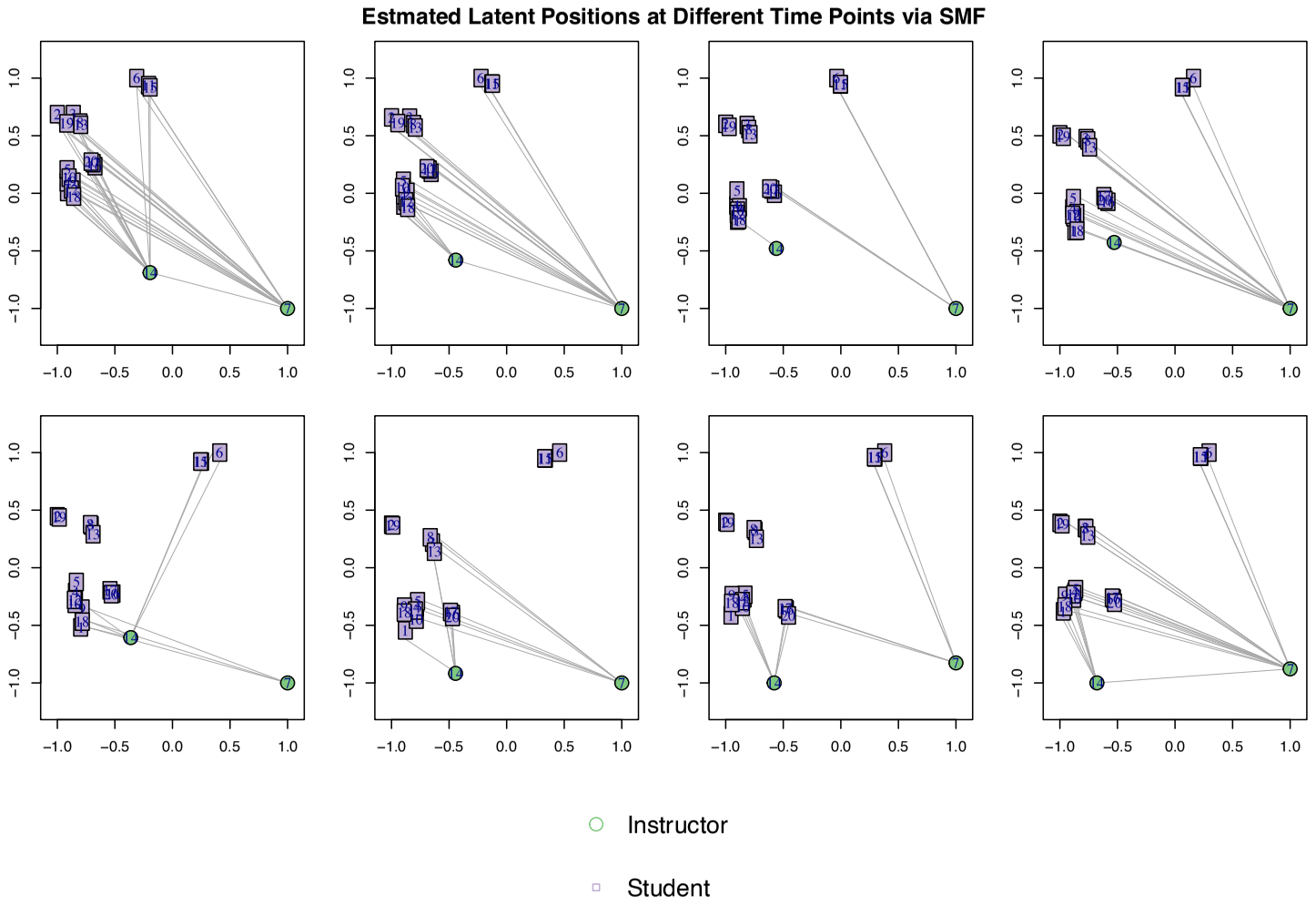

Figure 6: Dynamics of latent positions of all nodes from time point 1 to 8 for McFarland classroom data set via SMF. The locations of the nodes are the estimated latent positions. The top row of figures represents time points 1, 2, 3, and 4, while the bottom row represents time points 5, 6, 7, and 8. The edges between two nodes imply the interaction between the two observations within the corresponding time. Each number associated with the point is the index of the node.

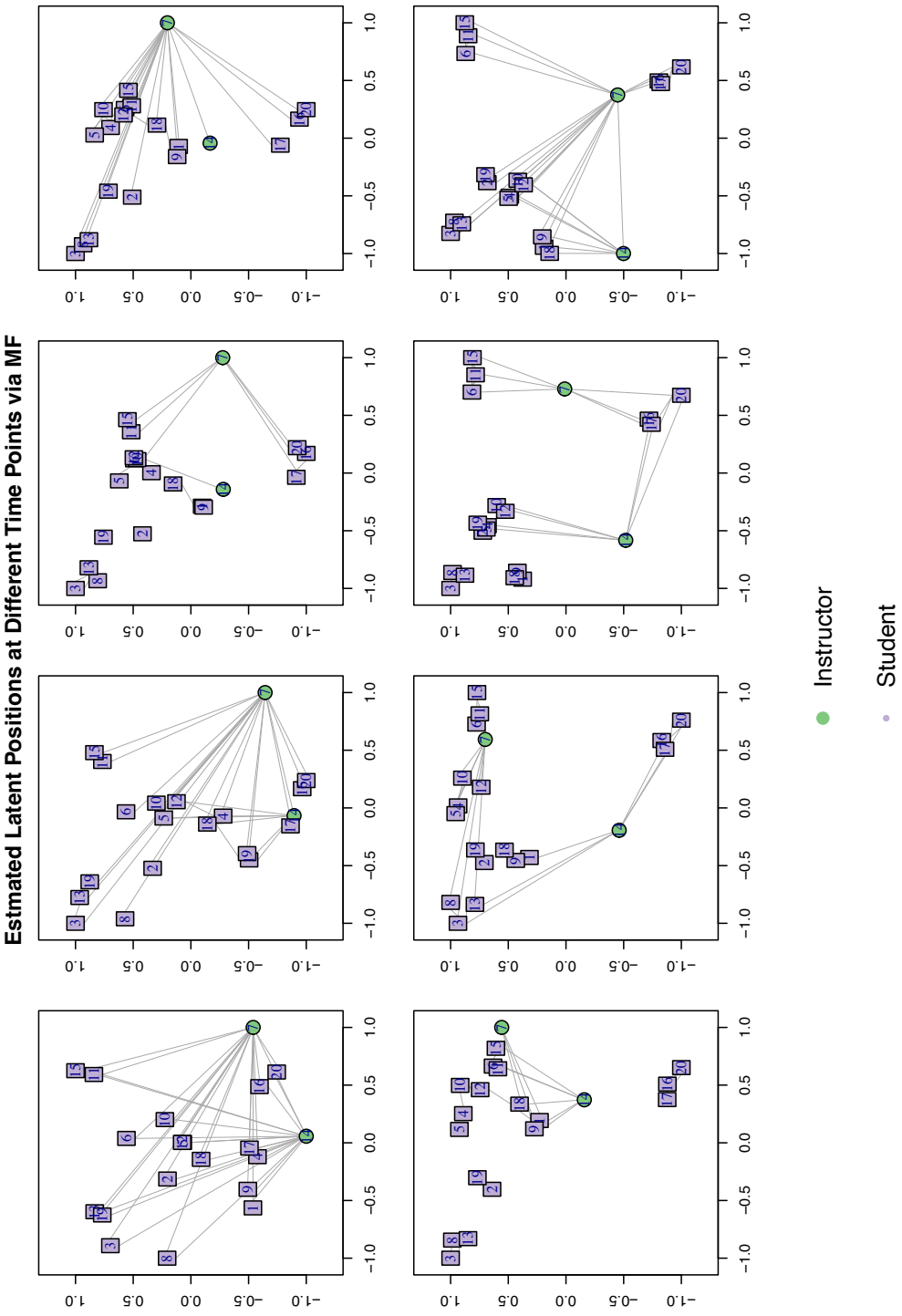

Figure 7: Dynamics of latent positions of all nodes from time point 1 to 8 for McFarland classroom data set via MF. The locations of the nodes are the estimated latent positions. The top row of figures represents time points 1, 2, 3, and 4, while the bottom row represents time points 5, 6, 7, and 8. The edges between two nodes imply the interaction between the two observations within the corresponding time. Each number associated with the point is the index of the node.

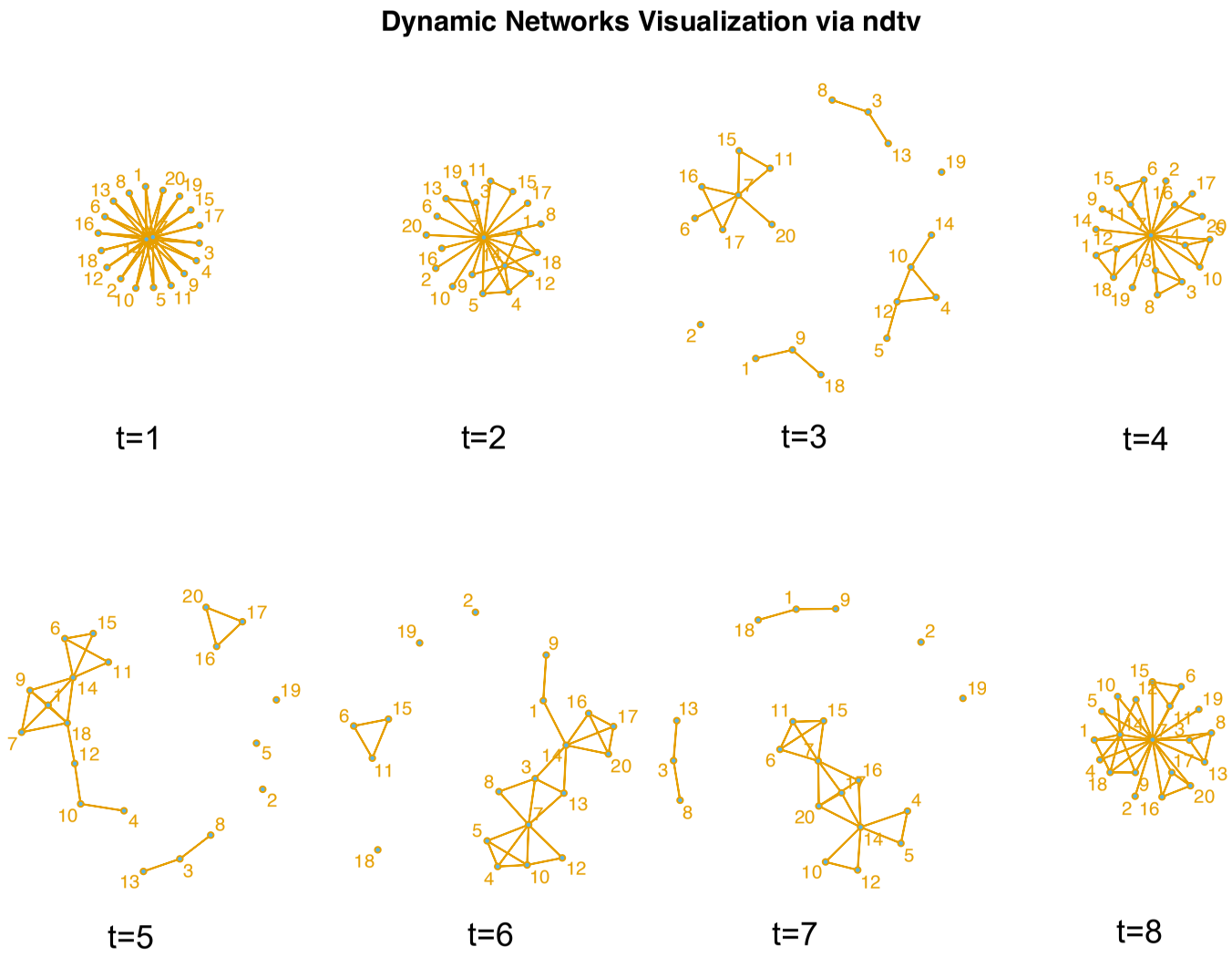

Figure 8: Visualization of the network dynamics from time point 1 to 8 for McFarland classroom data set using R package `ndtv`. From top left to top right, we have $t = 1, ..., 4$; from bottom left to bottom right, we have $t = 5, ..., 8$. Each number associated with the point is the index of the node.

Appendix A. Appendix

A.1 Proof of Theorem 2

Within the networks, we adopt the hypotheses constructions for some low-rank matrices, while among the networks, we adopt the test constructions similar to the constructions in total variational literature (Padilla et al., 2017).

For $\mathcal{U} = \{\boldsymbol{U}_t\}_{t=1}^T$ with $\boldsymbol{U}_t = [\boldsymbol{u}_{1t}, ..., \boldsymbol{u}_{nt}]'$ and $\mathcal{V} = \{\boldsymbol{V}_t\}_{t=1}^T$ with $\boldsymbol{V}_t = [\boldsymbol{v}_{1t}, ..., \boldsymbol{v}_{nt}]'$, let

$$d^2(\mathcal{U}, \mathcal{V}) = \sum_{t=1}^{T} \sum_{i \neq j=1}^{n} (\boldsymbol{u}_{it}'\boldsymbol{u}_{jt} - \boldsymbol{v}_{it}'\boldsymbol{v}_{jt})^2,$$

and

$$d_0^2(\boldsymbol{U}_t, \boldsymbol{V}_t) = \sum_{i \neq j=1}^{n} (\boldsymbol{u}_{it}'\boldsymbol{u}_{jt} - \boldsymbol{v}_{it}'\boldsymbol{v}_{jt})^2.$$

Hypothesis constructions for the low-rank part
First, we need the following lemma to obtain sparse Varshamov-Gilbert Bound under Hamming distance for the low-rank subset construction:

Lemma 9 (Lemma 4.10 in Massart, 2007) *Let $\Omega = \{0,1\}^n$ and $1 \leq s \leq n/4$. Then there exists a subset $\{w^{(1)}, ..., w^{(M)}\} \subset \Omega$ such that*

1. *$\|w^{(i)}\|_0 = s$ for all $1 \leq i \leq M$;*
2. *$\|w^{(i)} - w^{(j)}\|_0 \geq s/2$ for $0 \leq i \neq j \leq M$;*
3. *$\log M \geq cs \log(n/s)$ with $c \geq 0.233$.*

Let $\Omega_M = \{\boldsymbol{w}^{(1)}, ..., \boldsymbol{w}^{(M)}\} \subset \{0,1\}^{(n-d+1)/2}$ constructed based on the above Lemma (the construction holds under $n - d + 1 \geq 8$). For each $\boldsymbol{w}$, we can construct a $n \times d$ matrix as follows:

$$\boldsymbol{U}^w = \begin{bmatrix} \boldsymbol{v}^w & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{d-1} \end{bmatrix} \quad \text{with} \quad \boldsymbol{v}^w = \begin{bmatrix} 1 \\ ... \\ 1 \\ \epsilon \boldsymbol{w} \end{bmatrix} \in \mathbb{R}^{n-d+1}, w \in \Omega_M \tag{A.1}$$

where the first $(n-d+1)/2$ components for $\boldsymbol{v}^w$ are all ones.

The effect of this construction is that: for different $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \Omega_M$, since $s/2 \leq \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_0 \leq 2s$ and $\|\boldsymbol{U}^w\|_F \leq \sqrt{n}$, we have

$$\begin{aligned} d_0(\boldsymbol{U}^{w_1}, \boldsymbol{U}^{w_2}) \leq \|\boldsymbol{U}^{w_1}\boldsymbol{U}'^{w_1} - \boldsymbol{U}^{w_2}\boldsymbol{U}'^{w_2}\|_F &\leq \|\boldsymbol{U}^{w_1}(\boldsymbol{U}'^{w_1} - \boldsymbol{U}'^{w_2})\|_F + \|(\boldsymbol{U}^{w_1} - \boldsymbol{U}^{w_2})\boldsymbol{U}'^{w_2}\|_F \\ &\leq 2\sqrt{n}\|\boldsymbol{U}^{w_1} - \boldsymbol{U}^{w_2}\|_F = 2\sqrt{n}\|\boldsymbol{v}^{w_1} - \boldsymbol{v}^{w_2}\|_2 \leq 2\sqrt{2ns}\epsilon. \end{aligned}$$

In addition, consider $A := \{i + (n-d+1)/2 : w_{1i} \neq 0\}$, $B := \{j + (n-d+1)/2 : w_{2j} \neq 0\}$, $C := A \cap B$, where w_{1i}, w_{2j} are i and jth component of w_1 and w_2. We have $|C| \leq s/2$, $|A - C| \geq s/2$ and $|B - C| \geq s/2$. By direct calculation, we have

$$d_0^2(\boldsymbol{U}^{w_1}, \boldsymbol{U}^{w_2}) = \sum_{i \neq j}^{n-d+1} (v_i^{w_1} v_j^{w_1} - v_i^{w_2} v_j^{w_2})^2.$$

By only considering the sum for $i \in \{1, ..., (n-d+1)/2\}, j \in A-C$ where $v_i^{w_1} = v_i^{w_2} = 1$ $v_j^{w_1} = \epsilon$ and $v_j^{w_2} = 0$ and $i \neq j$, we have

$$d_0^2(\boldsymbol{U}^{w_1}, \boldsymbol{U}^{w_2}) \geq \sum_{i=1}^{(n-d+1)/2} \sum_{j \in A-C} (\epsilon w_{1j} - \epsilon w_{2j})^2 \geq \frac{s(n-d+1)}{4}\epsilon^2.$$

Hypothesis constructions for the total variational denoising part

As in the total variation denoising literature, we partition the set $\{1, ..., T\}$ into m groups $S_1, S_2, ..., S_m$ such that $S_1 = \{1, ..., k\}$, $S_2 = \{k+1, ..., 2k\}$, ..., $S_m = \{(m-1)k+1, ..., T\}$, where k will be decided later. Then we have $k(m-1)+1 \leq T \leq km$. For simplicity, we assume the partition is even $T = km$, otherwise we can consider $T' = km$, which has the same rate with T since $km > T > km - k + 1$. As in the literature in nonparametric regression, we need to obtain the optimal order of k or m.

Let $\boldsymbol{w}_0 = [0, ..., 0]' \in \mathbb{R}^{(n-d-1)/2}$ and

$$\boldsymbol{U}^0 = \begin{bmatrix} \boldsymbol{v}^{w_0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{d-1} \end{bmatrix} \quad \text{with} \quad \boldsymbol{v}^{w_0} = \begin{bmatrix} 1 \\ ... \\ 1 \\ \boldsymbol{w}_0 \end{bmatrix} \in \mathbb{R}^{n-d+1}$$

and $\mathcal{U}^0 = [\boldsymbol{U}^0, ..., \boldsymbol{U}^0]$. We need the Varshamov-Gilbert Bound 9 again to introduce another binary coding: let $\Omega_r = \{\boldsymbol{\phi}^{(1)}, ..., \boldsymbol{\phi}^{(M_0)}\} \subset \{0,1\}^m$, such that $\|\boldsymbol{\phi}^{(i)}\|_0 = s_0$ with $s_0 \leq m/4$ for all $1 \leq i \leq M_0$ and and $\|\boldsymbol{\phi}^{(i)} - \boldsymbol{\phi}^{(j)}\|_0 \geq s_0/2$ for $0 \leq i < j \leq M_0$ and $\log M_0 \geq cs_0 \log(m/s_0)$ with $c \geq 0.233$. The construction holds under $m \geq 4$.

Then the construction is based on a mixture of product space of Ω_M and group structure for $S_1, ...S_m$:

$$\begin{aligned} \Theta_\epsilon = \{\mathcal{X}_{...}^{(w,\phi)} : \\ & \boldsymbol{X}_t^{(w,\phi)} = \boldsymbol{U}^{w^{(i)}}, \quad \forall t \in S_j \quad \text{if} \quad \phi_j = 1, \\ & \boldsymbol{X}_t^{(w,\phi)} = \boldsymbol{U}^0, \quad \forall t \in S_j, \quad \text{if} \quad \phi_j = 0, \\ & \boldsymbol{w}^{(i)} \in \Omega_M, \forall i = 1, ..., s_0, \, \boldsymbol{w}^{(i)} \text{ are chosen with replacement}, \boldsymbol{\phi} \in \Omega_r\}, \end{aligned} \tag{A.2}$$

For example, when $\phi = (0,1,0,1,0,1,...)$, $\boldsymbol{X}_1^{(w,\phi)}$,..,$\boldsymbol{X}_T^{(w,\phi)}$ is:

$$\underbrace{\boldsymbol{U}^0, ..., \boldsymbol{U}^0}_{|S_1|}, \underbrace{\boldsymbol{U}^{w^{(1)}}, ..., \boldsymbol{U}^{w^{(1)}}}_{|S_2|}, \underbrace{\boldsymbol{U}^0, ..., \boldsymbol{U}^0}_{|S_3|}, \underbrace{\boldsymbol{U}^{w^{(2)}}, ..., \boldsymbol{U}^{w^{(2)}}}_{|S_4|}, ...$$

We have $|\Theta_\epsilon| = M_0 M^{s_0}$. In addition, for $\mathcal{U}, \mathcal{V} \in \Theta_\epsilon$, we have

$$d^2(\mathcal{U}, \mathcal{V}) = \sum_{t=1}^{T} d_0^2(\boldsymbol{U}_t, \boldsymbol{V}_t) \geq ks_0 \frac{s(n-d+1)}{8}\epsilon^2.$$

Besides, the KL divergence between any elements $\mathcal{U} \in \Theta_\epsilon$ and $\mathcal{U}^0$ can be upper bounded:

$$D_{KL}(P_{\mathcal{U}} \,||\, P_{\mathcal{U}^0}) \leq C_0 d^2(\mathcal{U}, \mathcal{V}) \leq 16 C_0 kns_0 s\epsilon^2, \tag{A.3}$$

for some constant $C_0 > 0$ ($C_0 = 1$ for the binary case, $C_0 = 1/(2\sigma^2)$ for the Gaussian case).

We use the following lemma to finally obtain the minimax lower bounds.

Lemma 10 (Theorem 2.5 in Tsybakov, 2008) *Suppose $M \geq 2$ and (Θ, d) contains elements $\theta_0, ..., \theta_M$ such that $d(\theta_i, \theta_j) \geq 2s > 0$ for any $0 \leq i \leq j \leq M$ and $\sum_{i=1}^{M} D_{KL}(P_{\theta_i}, P_0)/M \leq \alpha \log M$ with $0 < \alpha < 1/8$. Then we have*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}, \theta) \geq s) \geq \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right).$$

To adopt the above Lemma, it suffices to show

$$16 C_0 k s_0 s n \epsilon^2 \leq \alpha \log(M_0 M^{s_0}) = \alpha \log(|\Theta_\epsilon|),$$

with $\alpha < 1/8$. Let $s = (n-d-1)/8$, $s_0 = m/4$, according to lemma 9, it's enough to set

$$Tn(n-1)\epsilon^2 \leq \frac{cm \log 4}{4} + \frac{cm(n-d-1)\log 4}{16} = \frac{cm(n-d+7)\log 4}{16}, \tag{A.4}$$

with $c = 0.233/C_0$.

Minimax rate for point-wise dependence

Based on our construction, $2(m-1)s\epsilon \leq L$ should be satisfied, and we consider the following different cases:

Case 1: If there exist constants $c_0, c_0' > 0$ such that $c_0'/\sqrt{nT} < L \leq c_0(T-1)n^{1/2}$, which results in $16^3 TL^2/\{4sc \log 4\} < T(T-1)^2$, and $16^3 TL^2/(4c \log 4s) > 36 = 4 * (4-1)^2$. Therefore, since $n \geq 2d$, by assigning $\epsilon = L/\{2(m-1)s\}$ it is enough to let m satisfy

$$\frac{TL^2}{(m-1)^2} \leq \frac{4cms \log 4}{16^3},$$

which is

$$m(m-1)^2 \geq \frac{16^3 TL^2}{4c \log 4s}, \tag{A.5}$$

and m can be chosen within $4 \leq m \leq T$. Let m be the least integer such that the above inequality hold, then there exists a constant c_2, such that $m \leq c_2 T^{1/3} L^{2/3} n^{-1/3}$, which implies

$$ks_0 \frac{s(n-d+1)}{8}\epsilon^2 \gtrsim L^{\frac{2}{3}} n^{\frac{2}{3}} T^{\frac{1}{3}}.$$

Case 2: If there exists a constant $c_0 > 0$ such that $L > c_0(T-1)\sqrt{n}$, which results in $16^3 TL^2/\{cs \log 4\} > T(T-1)^2$, then we choose $m = T$, $\epsilon = c_0/\sqrt{n}$ such that $(m-1)s\epsilon \leq c_0(T-1)\sqrt{n} \leq L/2$. Then

$$ks_0 \frac{s(n-d+1)}{8} \gtrsim Tn.$$

Case 3: If $L < c_0'/\sqrt{nT}$ for some constant $c_0' > 0$ such that the least integer solution of inequality (A.5) satisfying $m < 4$. Then the above hypothesis construction in Equation (A.2) doesn't hold. Instead of considering the construction in Equation (A.2), we consider T copies of the same matrix, which implies the choice of m is 1. Note that the constraint on the norm of the difference of the matrix is automatically satisfied when all matrices are the same. By constructing the following subset

$$\Theta_\epsilon = \{\mathcal{X}^{(w)} : \boldsymbol{X}_t^{(w)} = \boldsymbol{U}^w, \quad \forall t = 1, ..., T, \boldsymbol{w} \in \Omega_M\}. \tag{A.6}$$

the KL divergence between any elements $\mathcal{U} \in \Theta_\epsilon$ and $\mathcal{U}^0$ can be upper bounded:

$$D_{KL}(P_{\mathcal{U}} \,||\, P_{\mathcal{U}^0}) \leq C_0 d_0^2(\boldsymbol{U}_t, \boldsymbol{V}_t) \leq 16 C_0 T s n \epsilon^2, \tag{A.7}$$

for some constant $C_0 > 0$. Then it suffices to let

$$16 T s n \epsilon^2 \leq \frac{c(n-d-1)\log 4}{16} \leq \alpha \log(M).$$

Therefore, based on the above equation, we need to choose $\epsilon = \sqrt{1/(nT)}$. Then we have

$$\frac{k s_0 s (n-d+1)\epsilon^2}{8} \gtrsim n.$$

Finally, based on Markov's inequality, by combining the above three cases, we have

$$\inf_{\hat{\mathcal{X}}} \sup_{\mathcal{X} \in \Theta_\epsilon} \mathbf{E}_{\mathcal{X}} \left[\frac{1}{Tn(n-1)} d^2(\hat{\mathcal{X}}, \mathcal{X}) \right] \gtrsim \min\left\{ \frac{L^{\frac{2}{3}}}{n^{\frac{4}{3}} T^{\frac{2}{3}}}, \frac{1}{n} \right\} + \frac{1}{nT}.$$

Therefore, the final conclusion holds.

A.2 Proof of Theorem 3

Proof As discussed in Bhattacharya et al. (2019), under the prior concentration condition that

$$\Pi(B_{n,T}(\mathcal{X}^*; \epsilon_n)) \geq e^{-Tn(n-1)\epsilon_{n,T}^2},$$

we can obtain the convergence of the α-divergence:

$$D_\alpha(p_{\mathcal{X}}, p_{\mathcal{X}^*}) = \frac{1}{\alpha - 1} \log \int (p_{\mathcal{X}^*})^\alpha (p_{\mathcal{X}})^{1-\alpha} d\mu.$$

Based on calculation, for Gaussian likelihood, we have $\max\{D_{KL}(p_{\mathcal{X}}, p_{\mathcal{X}^*}), V_2(p_{\mathcal{X}}, p_{\mathcal{X}^*})\} \lesssim \sum_{i \neq j, t} (\boldsymbol{x}_{it}' \boldsymbol{x}_{jt} - \boldsymbol{x}_{it}^{*'} \boldsymbol{x}_{jt}^*)^2$ where $V_2(p_{\mathcal{X}}, p_{\mathcal{X}^*})$ is the second moment of KL ball. For the Bernoulli likelihood, by Lemma 14, we have

$$D_{KL}(p_{\mathcal{X}}, p_{\mathcal{X}^*}) = \int p_{\mathcal{X}^*} \log(\frac{p_{\mathcal{X}^*}}{p_{\mathcal{X}}}) d\mu \leq \sum_{t=1}^{T} \sum_{i \neq j = 1}^{n} (\boldsymbol{x}_{it}' \boldsymbol{x}_{jt} - \boldsymbol{x}_{it}^{*'} \boldsymbol{x}_{jt}^*)^2.$$

Moreover, we have

$$V_2(p_{\mathcal{X}}, p_{\mathcal{X}^*}) := \int p_{\mathcal{X}^*} \log^2(\frac{p_{\mathcal{X}^*}}{p_{\mathcal{X}}}) d\mu \leq \sum_{i \neq j = 1}^{n} \sum_{t=1}^{T} 2 p_{x_{it}^*, x_{jt}^*} (\log \frac{p_{x_{it}^*, x_{jt}^*}}{p_{x_{it}, x_{jt}}})^2 + 2(1 - p_{x_{it}^*, x_{jt}^*}) (\log \frac{1 - p_{x_{it}^*, x_{jt}^*}}{1 - p_{x_{it}, x_{jt}}})^2. \tag{A.8}$$

Under the conditions that $p_{x_{it}^*, x_{jt}^*} := 1/\{1 + \exp(-\boldsymbol{x}_{it}^{*'} \boldsymbol{x}_{jt}^*)\}$ is bounded away from 0 and 1. The right hand side of Equation (A.8) is bounded above by $\sum_{i \neq j, t} (\boldsymbol{x}_{it}' \boldsymbol{x}_{jt} - \boldsymbol{x}_{it}^{*'} \boldsymbol{x}_{jt}^*)^2$ multiplied by some positive constant. Therefore, we also have $\max\{D_{KL}(p_{\mathcal{X}}, p_{\mathcal{X}^*}), V_2(p_{\mathcal{X}}, p_{\mathcal{X}^*})\} \lesssim \sum_{i \neq j, t} (\boldsymbol{x}_{it}' \boldsymbol{x}_{jt} - \boldsymbol{x}_{it}^{*'} \boldsymbol{x}_{jt}^*)^2$ for the binary case. Hence we only need to lower bound the prior

probability of the set $\{\sum_{i\neq j,t}(\boldsymbol{x}'_{it}\boldsymbol{x}_{jt}-\boldsymbol{x}^{*'}_{it}\boldsymbol{x}^*_{jt})^2 \leq n(n-1)T\epsilon^2\} \supset \{\max_t \max_{i\neq j}(\boldsymbol{x}'_{it}\boldsymbol{x}_{jt}-\boldsymbol{x}^{*'}_{it}\boldsymbol{x}^*_{jt})^2 \leq \epsilon^2\}$. Given $i\neq j, t$ we have

$$\begin{aligned}|\boldsymbol{x}'_{it}\boldsymbol{x}_{jt}-\boldsymbol{x}^{*'}_{it}\boldsymbol{x}^*_{jt}| &\leq |(\boldsymbol{x}'_{it}-\boldsymbol{x}^{*'}_{it})\boldsymbol{x}^*_{jt}| + |\boldsymbol{x}'_{it}(\boldsymbol{x}_{jt}-\boldsymbol{x}^*_{jt})| \\ &\leq \max_i \|\boldsymbol{x}_{it}-\boldsymbol{x}^*_{it}\|_2(\|\boldsymbol{x}_{it}-\boldsymbol{x}^*_{it}\|_2+2\|\boldsymbol{x}^*_{it}\|_2) \leq \max_i \|\boldsymbol{x}_{it}-\boldsymbol{x}^*_{it}\|_2(\|\boldsymbol{x}_{it}-\boldsymbol{x}^*_{it}\|_2+2C).\end{aligned}$$

Then when $\max_i \|\boldsymbol{x}_{it}-\boldsymbol{x}^*_{it}\|_2 \leq \epsilon/\{(2+c_0)C\} \leq C$ for some constants $c_0>1$, we have

$$\max_i \|\boldsymbol{x}_{it}-\boldsymbol{x}^*_{it}\|_2(2C+\|\boldsymbol{x}_{it}-\boldsymbol{x}^*_{it}\|_2) \leq \frac{\epsilon}{(2+c_0)C}3C \leq \epsilon.$$

Denote $E_0 = \max_i \|\boldsymbol{x}_{it}-\boldsymbol{x}^*_{it}\|_2 \leq \epsilon/\{(2+c_0)C\}$, $E_1 = \{\max_{i,j,t}|(X_{ijt}-X_{ij1})-(X^*_{ijt}-X^*_{ij1})| \leq \epsilon_0\}$, $E_2 = \{\max_{i,j}|X_{ij1}-X^*_{ij1}| \leq \epsilon_0\}$ with $\epsilon_0 = \epsilon/((2+c_0)C\sqrt{d})$. Then we have

$$\Pi(E_0) \geq \Pi(E_1)\,\Pi(E_2) = \prod_{i,j}\Pi\left(\sup_{t\geq 2}|\tilde{X}_{ijt}-\tilde{X}^*_{ijt}| \leq \epsilon_0\right)\prod_{i,j}\Pi\left(|X_{ij1}-X^*_{ij1}| \leq \epsilon_0\right),$$

where $\tilde{X}_{ijt} = X_{ijt}-X_{ij1}$ for all i,j,t.

Given i,j, $\xi_{ijt} \sim \mathcal{N}(0,\tau^2)$ for $t\geq 2$, we can denote $\tilde{X}_{ijt} = \sum_{s=1}^t \xi_{ijs}$ and $(\tilde{X}_{ij2},...,\tilde{X}_{ijt})' \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}_0)$. Denote $\tilde{\boldsymbol{x}}^{*'}_{ij} = (\tilde{X}^*_{ij2},...,\tilde{X}^*_{ijt})'$. Based on multivariate Gaussian concentration through Anderson's inequality, we have

$$\begin{aligned}\Pi(E_1) &\geq \prod_{i,j} P\left(\sup_{t\geq 2}|\tilde{X}_{ijt}-\tilde{X}^*_{ijt}| \leq \epsilon_0\right) \\ &\geq \prod_{i,j}\exp(-\frac{\tilde{\boldsymbol{x}}^{*'}_{ij}\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{x}}^*_{ij}}{2})\Pi\left(\sup_t|\tilde{X}_{ijt}| \leq \epsilon_0\right).\end{aligned} \tag{A.9}$$

By the definition of $\boldsymbol{\Sigma}_0$, we have

$$-\frac{\tilde{\boldsymbol{x}}^{*'}_{ij}\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{x}}_{ij}{}^*}{2} = -\sum_{t=2}^T\frac{(\tilde{X}^*_{ijt}-\tilde{X}^*_{ij(t-1)})^2}{2\tau^2} = -\sum_{t=2}^T\frac{(X^*_{ijt}-X^*_{ij(t-1)})^2}{2\tau^2},$$

where $\tilde{X}^*_{ij1}=0$. For the second factor in Equation (A.9) , given i,j, we consider a Gaussian process $\{\tilde{X}_{ij}(s), 0\leq s\leq 1\}$ induced by $(\tilde{X}_{ij2},...,\tilde{X}_{ijt})$ such that $\tilde{X}_{ij}((s-2)/(T-2)) = \tilde{X}_{ijt}$, $\tilde{X}_{ij}(0)=0$ and all other values are obtained through interpolations: $\tilde{X}_{ij}(s) = w_0\tilde{X}_{ij(t-1)} + (1-w_0)\tilde{X}_{ijt}$ $\forall w_0\in(0,1)$ with $s = w_0(t-3)/(T-2)+(1-w_0)(t-2)/(T-2)$. Then clearly, we have

$$\Pi\left(\sup_{t=2,...,T}|\tilde{X}_{ijt}| \leq \delta\right) \geq \Pi\left(\sup_{s\in[0,1]}|\tilde{X}_{ij}(s)| \leq \delta\right),$$

for any $\delta>0$. Denote $\sigma^2(h) = E(\tilde{X}(s+h)-\tilde{X}(s))^2 = hT\tau^2$. Then $\sigma^2(h)$ is linear in h hence concave. In addition, $\sigma(h)/(h^{1/2}) = \sqrt{T\tau^2}$, which is non-decreasing in $(0,1)$. Based on Lemma 13, we have

$$\Pi(\sup_{0\leq s\leq 1}|\tilde{X}(s)| \leq \delta) \geq C_4\exp(-C_3\frac{T\tau^2}{\delta^2})$$

for $\delta > 0$ with constants $C_3, C_4 > 0$.

Therefore, for some constant $C_3, C_4 > 0$, we have

$$\begin{aligned}\Pi(E_1) &\geq C_4 \exp\left[-\sum_{t=2}^{T} \frac{\|\boldsymbol{X}_t^* - \boldsymbol{X}_{t-1}^*\|_F^2}{2\tau^2} - C_3 \frac{nT\tau^2}{\epsilon^2}\right] \\ &\geq C_4 \exp\left[-n\sum_{t=2}^{T} \max_i \frac{\|\boldsymbol{x}_{it}^* - \boldsymbol{x}_{i(t-1)}^*\|_2^2}{2\tau^2} - C_3 \frac{nT\tau^2}{\epsilon^2}\right].\end{aligned} \tag{A.10}$$

For PWD, with condition (9), we have

$$\Pi(E_1) \geq C_4 \exp\left[-\frac{-C_0^2 L^2}{2nT\tau^2} - C_3 \frac{Tn\tau^2}{\epsilon^2}\right].$$

Moreover, by taking that

$$\tau^2 = \epsilon L/(nT)$$

we can obtain

$$\log \Pi(E_1) \gtrsim -\frac{L}{\epsilon}.$$

For the initial error concentration $\Pi(E_2)$, by the mean-zero Gaussian of X_{ij1} for all i, j, we have the concentration:

$$\begin{aligned}\Pi(E_2) = \prod_{i,j} \Pi\left(|X_{ij1} - X_{ij1}^*| \leq \epsilon_0\right) &\geq \frac{1}{(\sqrt{2\pi}\sigma_0)^{nd}} \exp(-\sum_{i,j} \frac{X_{ij1}^{*2}}{2\sigma_0^2})(2\epsilon)^{nd} \\ &\gtrsim \exp\left[-\frac{\|\boldsymbol{X}_1^*\|_F^2}{2\sigma_0^2} - nd - nd\log(\frac{1}{\epsilon})\right].\end{aligned}$$

Note that σ is a constant and $\|\boldsymbol{X}_1^*\|_F^2 = O(n)$. We have $\log \Pi(E_2) \gtrsim -n\log(1/\epsilon)$.

Then the rate $\epsilon_{n,T} = L^{1/3}T^{-1/3}n^{-2/3} + \sqrt{\log(nT)/nT}$ can be obtained by letting the smallest possible $\epsilon_{n,T}$ such that $n(n-1)T\epsilon_{n,T}^2 \gtrsim \max\{L/\epsilon_{n,T}, n\log(1/\epsilon_{n,T})\}$.

Finally, this additive rate helps in the choice of the transition τ. In particular, when $L < \log^{3/2}(nT)\sqrt{n/T}$ such that $L^{1/3}T^{-1/3}n^{-2/3} \lesssim \sqrt{\log(nT)/nT}$, the choice of τ can be relaxed as long as $-\log \Pi(E_1) \lesssim n\log(nT)$. Therefore, let $\tau^2 = \log^2(nT)/(nT^2)$ in this case, we have

$$Tn\tau^2/\epsilon_{n,T}^2 \lesssim n\log(nT), \quad \text{and} \quad L^2/(nT\tau^2) = n\log(nT). \tag{A.11}$$

Therefore, the final choice of τ that guarantees the optimal convergence rate satisfies $\tau^2 = \log^2(nT)/(nT^2) + \epsilon_{n,T}L/(nT)$.

By Theorem 3.1 in Bhattacharya et al. (2019), the prior concentration $\Pi(B_{n,T}(\mathcal{X}^*; \epsilon_{n,T})) \geq \exp(-Tn(n-1)\epsilon_{n,T}^2)$ implies that the posterior contraction of the averaged α-divergence for any $0 < \alpha < 1$ is at the rate $\epsilon_{n,T}^2$. For the Gaussian case, by the direct calculation (Gil et al., 2013), we can obtain that the α-divergence is lower bounded by the squared loss function up to some constant factor when the variance of the likelihood is fixed. For binary case, based on the boundness of the truth and Lemma 16, which indicates that the 1/2 divergence is lower bounded by the squared loss function up to some constant factor, we can achieve the results in equation (12).

■

A.3 Proof of Theorem 4

Proof Let $\sigma_0^{*2} = 1$ and $\tau^{*2} = \epsilon_{n,T}L/(nT) + \log^2(nT)/(nT^2)$. In the proof of Theorem 3, we show the prior concentration conditional on $\sigma_0 = c_1\sigma_0^*$ and $\tau = c_2\tau^*$ for any constants $c_1, c_2 > 0$ is sufficient:

$$-\log\{\Pi(B_{n,T}(\mathcal{X}^*;\epsilon_{n,T}) \mid c_1\sigma_0^*, c_2\tau^*)\} \lesssim Tn(n-1)\epsilon_{n,T}^2.$$

Therefore, by limiting on the subset $N(\sigma_0^*,\tau^*) = \{|\sigma_0^2 - \sigma_0^{2*}| \le \sigma_0^{2*}/2,\ |\tau^2 - \tau^{2*}| \le \tau^{2*}/2\}$, we have $-\log\Pi(B_{n,T}(\mathcal{X}^*;\epsilon_n) \mid \sigma_0,\tau) \lesssim Tn(n-1)\epsilon_{n,T}^2$. Then

$$\begin{aligned}&\int_{N(\sigma_0^*,\tau^*)} \Pi(B_{n,T}(\mathcal{X}^*;\epsilon_n) \mid \sigma_0,\tau)p(\tau)p(\sigma_0)d\tau\sigma_0\\ \gtrsim\ & P(|\tau^2-\tau^{*2}| \le \tau^{*2}/2)P(|\sigma_0^2-\sigma_0^{*2}| \le \sigma_0^{*2}/2)\exp(-Tn(n-1)c_0\epsilon_{n,T}^2),\end{aligned}$$

for some constant $c_0 > 0$. For σ_0, with the Inverse-gamma$(a_{\sigma_0}, b_{\sigma_0})$ prior where $a_{\sigma_0}, b_{\sigma_0}$ are constants, we have

$$P(|\sigma_0^2 - \sigma_0^{*2}| \le \sigma_0^{*2}/2) = P(1/2 \le \sigma_0^2 \le 3/2),$$

which is a fixed constant. For τ, with the Gamma(c_τ, d_τ) prior where c_τ, d_τ are constants, we have

$$\begin{aligned}P(|\tau^2-\tau^{*2}| \le \tau^{*2}/2) &= \int_{\epsilon_{n,T}L/(2nT)+\log^2(nT)/(2nT^2)}^{3\epsilon_{n,T}L/(2nT)+3\log^2(nT)/(2nT^2)} f_{c_\tau,d_\tau}(\tau^2)d\tau^2\\ &\ge \min_{|\tau^2-\tau^{*2}|\le\tau^{*2}/2} f_{c_\tau,d_\tau}(\tau^2)\left\{\epsilon_{n,T}L/(2nT) + \log^2(nT)/(2nT^2)\right\},\end{aligned}$$

where $f_{c_\tau,d_\tau}(\tau^2)$ is the density function of Gamma(c_τ, d_τ) prior. When $|\tau^2 - \tau^{*2}| \le \tau^{*2}/2$, we have

$$-\log\{\min_{|\tau^2-\tau^{*2}|\le\tau^{*2}/2} f_{c_\tau,d_\tau}(\tau^2)\} \lesssim \tau^{*2} - \log(\tau^{*2}) \lesssim \epsilon_{n,T}L/(nT) + \log^2(nT)/(4nT^2) + \log(nT).$$

Note that $\epsilon_{n,T} = o(1)$ due to $L = o(n^2T)$, we have

$$\epsilon_{n,T}L/(nT) \lesssim L/(nT) \lesssim n \lesssim n\log(nT) \lesssim n(n-1)T\epsilon_{n,T}^2.$$

In addition, $\log^2(nT)/(4nT^2) + \log(nT) \lesssim n(n-1)T\epsilon_{n,T}^2$ holds.

Moreover, we also have $-\log(\epsilon_{n,T}L/(nT) + \log^2(nT)/(2nT^2)) \lesssim \log(nT) \lesssim n(n-1)T\epsilon_{n,T}^2$. Therefore, we showed that under the prior Π, it holds that $-\log\Pi(|\tau^2 - \tau^{*2}| \le \tau^{*2}/2) \lesssim Tn(n-1)\epsilon_{n,T}^2$. Hence we have

$$\Pi(B_{n,T}(\mathcal{X}^*;\epsilon_n)) \ge \exp(-Tn(n-1)M\epsilon_{n,T}^2)$$

for large enough constant $M > 0$. With the choice $\epsilon_{n,T}^{new} = \sqrt{M}\epsilon_{n,T}$, we showed that the prior concentration is sufficient enough, and the rest of the proof is similar with Theorem 3 by applying Theorem 3.1 in Bhattacharya et al. (2019). ■

A.4 Proof of Proposition 5

Suppose $q(\beta)$, $q(\tau)$, $q(\sigma_0)$ and $q(\boldsymbol{x}_{j\cdot}), j \neq i$ are given. By the definition of ELBO and Equation (18), we have

$$\begin{aligned}\text{ELBO} = &\sum_{t=1}^{T-1} \int q_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)}) \log \psi_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)}) \mathrm{d}\mathcal{X} \\ &+ \int q_{it}(\boldsymbol{x}_{it}) \log \phi_{it}(\boldsymbol{x}_{it}) \mathrm{d}\mathcal{X} - \sum_{t=1}^{T} \int q_{it}(\boldsymbol{x}_{it}) \log q_{it}(\boldsymbol{x}_{it}) \mathrm{d}\mathcal{X} \\ &- \sum_{t=1}^{T-1} \int q_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)}) \{\log q_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)}) \\ &- \log q_{it}(\boldsymbol{x}_{it}) - \log q_{i(t+1)}(\boldsymbol{x}_{i(t+1)})\} \mathrm{d}\mathcal{X} + \text{other term}.\end{aligned}$$

By introducing Lagrange multiplier $\lambda_{it,i(t+1)}(\boldsymbol{x}_{i(t+1)})$ and $\lambda_{it,i(t-1)}(\boldsymbol{x}_{i(t-1)})$ for the marginalization conditions, for the term related with $q_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)})$, we have:

$$\log \psi_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)}) - \log \frac{q_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)})}{q_{it}(\boldsymbol{x}_{it}) q_{i(t+1)}(\boldsymbol{x}_{i(t+1)})} - \lambda_{it,i(t+1)}(\boldsymbol{x}_{i(t+1)}) - \lambda_{i(t+1),it}(\boldsymbol{x}_{it}) + \text{constant} = 0.$$

For the term related to $q_{it}(\boldsymbol{x}_{it})$, we have:

$$\log \phi_{it}(\boldsymbol{x}_{it}) - \log q_{it}(\boldsymbol{x}_{it}) + \lambda_{i(t+1),it}(\boldsymbol{x}_{it}) + \lambda_{i(t-1),it}(\boldsymbol{x}_{it}) + \text{constant} = 0.$$

Then by combining the above result, we have:

$$q_{it}(\boldsymbol{x}_{it}) \propto \phi_{it}(\boldsymbol{x}_{it}) \exp(\lambda_{i(t+1),it}(\boldsymbol{x}_{it}) + \lambda_{i(t-1),it}(\boldsymbol{x}_{it})). \tag{A.12}$$

Moreover, we have

$$\begin{aligned}q_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)}) \propto &\phi_{it}(\boldsymbol{x}_{it}) \phi_{i(t+1)}(\boldsymbol{x}_{i(t+1)}) \psi_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)}) \\ &\cdot \exp(\lambda_{i(t+2),i(t+1)}(\boldsymbol{x}_{i(t+1)}) + \lambda_{i(t-1),it}(\boldsymbol{x}_{it})).\end{aligned} \tag{A.13}$$

Finally, based on the marginalization property of $q_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)})$, we have the backward updating:

$$\exp(\lambda_{i(t+1),it}(\boldsymbol{x}_{it})) \propto \int \phi_{i(t+1)}(\boldsymbol{x}_{i(t+1)}) \psi_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)}) \exp(\lambda_{i(t+2),i(t+1)}(\boldsymbol{x}_{i(t+1)})) \mathrm{d}\boldsymbol{x}_{i(t+1)}$$

and forward updating:

$$\exp(\lambda_{it,i(t+1)}(\boldsymbol{x}_{i(t+1)})) \propto \int \phi_{it}(\boldsymbol{x}_{it}) \psi_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)}) \exp(\lambda_{i(t-1),it}(\boldsymbol{x}_{it})) \mathrm{d}\boldsymbol{x}_{it}.$$

Let $m_{it,i(t+1)}(\boldsymbol{x}_{i(t+1)}) = \exp(\lambda_{it,i(t+1)}(\boldsymbol{x}_{i(t+1)}))$ and $m_{it,i(t-1)}(\boldsymbol{x}_{i(t-1)}) = \exp(\lambda_{it,i(t-1)}(\boldsymbol{x}_{i(t-1)}))$. The Equation (22) directly follows Equation (A.12) and (A.13) after plugging in the forward and backward messages and therefore the proposition is proved.

A.5 Proof of Theorem 7

Proof The proof is based on Theorem 3.3 in Yang et al. (2020), where we need to provide upper bounds for

$$-\int \log \frac{P(\mathcal{Y} \mid \mathcal{X})}{P(\mathcal{Y} \mid \mathcal{X}^*)} q(\mathcal{X}) d\mathcal{X}$$

and

$$D_{KL}(q(\mathcal{X}) \,||\, p(\mathcal{X})),$$

where $q(\mathcal{X})$ is a variational distribution in the SMF family and $p(\mathcal{X})$ is the prior. Based on the definition of E_0 in the proof in subsection A.2, we have

$$B_{n,T}(\mathcal{X}^*; \epsilon) \supset E_0 := \{\max_{i,t} \|\boldsymbol{x}_{it} - \boldsymbol{x}_{it}^*\|_2 \leq \epsilon_0\},$$

with $\epsilon_0 = c_1 \epsilon_{n,T}$, for constant $c_1 > 0$. The above constraint can be written in a separate form:

$$E_0 = \cap_{i,t} \{\|\boldsymbol{x}_{it} - \boldsymbol{x}_{it}^*\|_2 \leq \epsilon_0\}.$$

Then we can choose $q(\mathcal{X})$ in the following way:

$$q(\mathcal{X}) \propto \prod_{i=1}^{n} \prod_{t=2}^{T} p(\boldsymbol{x}_{it} \mid \boldsymbol{x}_{i(t-1)}) \mathbb{1}\{\|\boldsymbol{x}_{it} - \boldsymbol{x}_{it}^*\|_2 \leq \epsilon_0\} \prod_{i=1}^{n} p(\boldsymbol{x}_{i1}) \mathbb{1}\{\|\boldsymbol{x}_{i1} - \boldsymbol{x}_{i1}^*\|_2 \leq \epsilon_0\},$$

where $p(\boldsymbol{x}_{it} \mid \boldsymbol{x}_{i(t-1)})$ and $p(\boldsymbol{x}_{i1})$ are components of priors. Note that the above variational distribution belongs to the SMF distribution family. We prove the above two bounds based on the current construction of $q(\mathcal{X})$. First, by Fubini's theorem and the definition of the prior, we have

$$\begin{aligned}
&\mathbf{E}_{\mathcal{X}^*} \left[-\int_{\mathcal{X}} q(\mathcal{X}) \log \frac{P(\mathcal{Y} \mid \mathcal{X})}{P(\mathcal{Y} \mid \mathcal{X}^*)} d\mathcal{X} \right] \\
&= \int_{\mathcal{X}} - \mathbf{E}_{\mathcal{X}^*} \left[\log \frac{P(\mathcal{Y} \mid \mathcal{X})}{P(\mathcal{Y} \mid \mathcal{X}^*)} \right] q(\mathcal{X}) d\mathcal{X} \\
&\leq \int_{B_n(\mathcal{X}^*, \epsilon)} D_{KL} \left[P(\mathcal{Y} \mid \mathcal{X}^*) \,||\, P(\mathcal{Y} \mid \mathcal{X}) \right] q(\mathcal{X}) d\mathcal{X} \leq n(n-1) T \epsilon^2.
\end{aligned}$$

Similarly, for the variance, by Jensen's inequality and Fubini's theorem, we have

$$\begin{aligned}
&\mathrm{Var}_{\mathcal{X}^*} \left[\int_{\mathcal{X}} q(\mathcal{X}) \log \frac{P(\mathcal{Y} \mid \mathcal{X})}{P(\mathcal{Y} \mid \mathcal{X}^*)} d\mathcal{X} \right] \\
&\leq \mathbf{E}_{\mathcal{X}^*} \left[\int_{\mathcal{X}} q(\mathcal{X}) \log \frac{P(\mathcal{Y} \mid \mathcal{X})}{P(\mathcal{Y} \mid \mathcal{X}^*)} d\mathcal{X} \right]^2 \\
&\leq \int_{B_n(\mathcal{X}^*, \epsilon)} V_2 \left[P(\mathcal{Y} \mid \mathcal{X}^*) \,||\, P(\mathcal{Y} \mid \mathcal{X}) \right] q(\mathcal{X}) d\mathcal{X} \leq n(n-1) T \epsilon^2.
\end{aligned}$$

Therefore, by Chebyshev's inequality, for any $D > 1$, based on the first and second moments of the above bounds, we have

$$\begin{aligned}
&P_{\mathcal{X}^*}\left[\int_{\mathcal{X}} q(\mathcal{X}) \log \frac{P(\mathcal{Y} \mid \mathcal{X})}{P\left(\mathcal{Y} \mid \mathcal{X}^*\right)} d\mathcal{X} \leq -Dn(n-1)T\epsilon^2\right] \\
&\leq P_{\mathcal{X}^*}\left[\int_{\mathcal{X}} q(\mathcal{X}) \log \frac{P(\mathcal{Y} \mid \mathcal{X})}{P\left(\mathcal{Y} \mid \mathcal{X}^*\right)} d\mathcal{X}\right. \\
&\left.- \mathbf{E}\left\{\int_{\mathcal{X}} q(\mathcal{X}) \log \frac{P(\mathcal{Y} \mid \mathcal{X})}{P\left(\mathcal{Y} \mid \mathcal{X}^*\right)} d\mathcal{X}\right\} \leq -(D-1)n(n-1)T\epsilon^2\right] \\
&\leq \mathrm{Var}_{\mathcal{X}^*}\left[\int_{\mathcal{X}} q(\mathcal{X}) \log \frac{P(\mathcal{Y} \mid \mathcal{X})}{P\left(\mathcal{Y} \mid \mathcal{X}^*\right)} d\mathcal{X}\right] / \left((D-1)^2 n^2 (n-1)^2 T^2 \epsilon^4\right) \\
&\leq \frac{4}{(D-1)^2 n(n-1)T\epsilon^2}
\end{aligned}$$

holds with probability $1 - 1/\{(D-1)^2 n(n-1)T\epsilon^2\}$.

This proves that when $n(n-1)T\epsilon \to \infty$, we have

$$-\int \log \frac{P(\mathcal{Y} \mid \mathcal{X})}{P(\mathcal{Y} \mid \mathcal{X}^*)} q(\mathcal{X}) d\mathcal{X} \leq Dn(n-1)T\epsilon^2$$

with probability converging to one.

In addition, based on the construction of the variational family, we have

$$D_{KL}(q(\mathcal{X}) || p(\mathcal{X})) = -\log(\Pi(E_0)),$$

since for any probability measure μ and measurable set A with $\mu(A) > 0$, we have $D_{KL}(\mu(\cdot \cap A)/\mu(A) \,||\, \mu) = -\log(\mu(A))$. By the proof in subsection A.2, we have $-\log(\Pi(E_0)) \lesssim -\log(\Pi(E_1 \cap E_2)) \lesssim \max\{L/\epsilon, n\log(1/\epsilon)\}$ for PWD(L) with Lipschitz condition. Therefore, the convergence of the α-divergence follows by Theorem 3.3 in Yang et al. (2020). Finally, the α-divergence is lower bounded by the loss according to the final part of the proof of Theorem 3.

■

A.6 Proof of Theorem 8

Proof Note that the prior now satisfies $p(\mathcal{X}, \tau, \sigma_0) = p(\mathcal{X} \mid \tau, \sigma_0)p(\tau)p(\sigma_0)$ and the variational distribution instead satisfies $q(\mathcal{X}, \tau, \sigma_0) = \prod_{i=1}^n q_i(\boldsymbol{x}_{i\cdot})q(\tau)q(\sigma_0)$. Let $\sigma_0^{*2} = 1$ and $\tau^{*2} = \epsilon_{n,T}L/(nT) + \log^2(nT)/(nT^2)$, we consider the following variational distribution:

$$\begin{aligned}
q(\mathcal{X}, \tau, \sigma_0) \propto &\prod_{i=1}^n \prod_{t=2}^T p(\boldsymbol{x}_{it} \mid \boldsymbol{x}_{i(t-1)}, \tau^*) \mathbb{1}\{\|\boldsymbol{x}_{it} - \boldsymbol{x}_{it}^*\|_2 \leq c_1\epsilon_{n,T}\} \\
&\times \prod_{i=1}^n p(\boldsymbol{x}_{i1} \mid \sigma_0^*) \mathbb{1}\{\|\boldsymbol{x}_{i1} - \boldsymbol{x}_{i1}^*\|_2 \leq c_1\epsilon_{n,T}\} \\
\times p(\tau)\mathbb{1}\{\tau^{*2} < \tau^2 < \tau^{*2} e^{\epsilon_{n,T}^2}\} &p(\sigma_0)\mathbb{1}\{\sigma_0^{*2} < \sigma_0^2 < \sigma_0^{*2} e^{\epsilon_{n,T}^2}\},
\end{aligned} \tag{A.14}$$

where c_1 is the constant used in the proof of Theorem 7. Given the prior, we still check the conditions

$$-\int \log \frac{P(\mathcal{Y} \mid \mathcal{X})}{P(\mathcal{Y} \mid \mathcal{X}^*)} q(\mathcal{X}, \tau, \sigma_0) d\mathcal{X} d\tau d\sigma_0 \lesssim Tn(n-1)\epsilon_{n,T}^2 \tag{A.15}$$

$$D_{KL}(q(\mathcal{X}, \tau, \sigma_0) \,||\, p(\mathcal{X}, \tau, \sigma_0)) \lesssim Tn(n-1)\epsilon_{n,T}^2 \tag{A.16}$$

First, the condition (A.15) directly follows the proof of Theorem 7 given the MF structure $q(\mathcal{X}, \tau, \sigma_0) = q(\mathcal{X})q(\tau)q(\sigma_0)$.

Then by the chain rule of KL divergence, we have

$$\begin{aligned} D_{KL}(q(\mathcal{X}, \tau, \sigma_0) \,||\, p(\mathcal{X}, \tau, \sigma_0)) = D_{KL}(q(\tau)||p(\tau)) + D_{KL}(q(\sigma_0)||p(\sigma_0)) \\ + \int q(\tau)q(\sigma_0) \int q(\mathcal{X}) \log \frac{q(\mathcal{X})}{p(\mathcal{X}|\tau, \sigma_0)} d\mathcal{X} d\tau d\sigma_0. \end{aligned} \tag{A.17}$$

With the Gamma(c_τ, d_τ) prior and $\epsilon_{n,T} < 1$, we have

$$\begin{aligned} D_{KL}(q(\tau)||p(\tau)) &= -\log(P(\tau^{*2} < \tau^2 < \tau^{*2} e^{\epsilon_{n,T}^2})) \\ &\leq -\log(\min_{\tau^{*2} < \tau^2 < \tau^{*2} e^{\epsilon_{n,T}^2}} f_{c_\tau, d_\tau}(\tau^2)(e^{\epsilon_{n,T}^2} - 1)) \\ &\overset{(i)}{\leq} -\log(\epsilon_{n,T}^2) - \log(\min_{\tau^{*2} < \tau^2 < \tau^{*2} e^{\epsilon_{n,T}^2}} f_{c_\tau, d_\tau}(\tau^2)) \\ &\overset{(ii)}{\lesssim} Tn(n-1)\epsilon_{n,T}^2 - \log(\min_{\tau^{*2} < \tau^2 < \tau^{*2} e^{\epsilon_{n,T}^2}} f_{c_\tau, d_\tau}(\tau^2)), \end{aligned} \tag{A.18}$$

where in (i) we use $e^x - 1 \geq x$ for any x and (ii) is because $\epsilon_{n,T}^2 \geq \log(nT)/(nT)$. In addition, by a similar approach with proof in Theorem 4, we have

$$-\log(\min_{\tau^{*2} < \tau^2 < \tau^{*2} e^{\epsilon_{n,T}^2}} f_{c_\tau, d_\tau}(\tau^2)) \lesssim \tau^{*2} - \log(\tau^{*2}) \lesssim n(n-1)T\epsilon_{n,T}^2.$$

With $\epsilon_{n,T} < 1$, we have $1 < \sigma_0^2 < e$ in the constrained region, where the density of Inverse-Gamma$(a_{\sigma_0}, b_{\sigma_0})$ is lower bounded by a constant. Hence,

$$D_{KL}(q(\sigma_0)||p(\sigma_0)) = -\log(P(\sigma_0^{*2} < \sigma_0^2 < \sigma_0^{*2} e^{\epsilon_{n,T}^2})) \lesssim -\log(\epsilon_{n,T}^2) \overset{(i)}{\lesssim} Tn(n-1)\epsilon_{n,T}^2, \tag{A.19}$$

where (i) is due to $\epsilon_{n,T}^2 \geq \log(nT)/(nT)$. For the third term of the KL divergence, we have

$$\int q(\mathcal{X}) \log \frac{q(\mathcal{X})}{p(\mathcal{X}|\tau, \sigma_0)} d\mathcal{X} = \int_{E_0} q(\mathcal{X}) \log \frac{p(\mathcal{X}|\tau^*, \sigma_0^*)}{p(\mathcal{X}|\tau, \sigma_0)} d\mathcal{X} - \log(\Pi(E_0 \mid \tau^*, \sigma_0^*)).$$

Note that we already have $-\log(\Pi(E_0 \mid \tau^*, \sigma_0^*)) \lesssim Tn(n-1)\epsilon_{n,T}^2$ by the proof of the prior concentration in subsection A.2.

Moreover, we have the density,

$$p(\mathcal{X} \mid \tau, \sigma_0) = \frac{1}{(\sqrt{2\pi})^{nTd}} \exp\left\{ -\frac{n(T-1)d}{2} \log(\tau^2) - \frac{nd}{2} \log(\sigma_0^2) - \frac{\|\boldsymbol{X}_1\|_F^2}{2\sigma_0^2} - \frac{\sum_{t=2}^T \|\boldsymbol{X}_t - \boldsymbol{X}_{t-1}\|_F^2}{2\tau^2} \right\},$$

which implies that

$$\log\frac{p(\mathcal{X}|\tau^*,\sigma_0^*)}{p(\mathcal{X}|\tau,\sigma_0)} = \frac{n(T-1)d}{2}\log(\tau^2) - \frac{n(T-1)d}{2}\log(\tau^{*2}) + \frac{nd}{2}\log(\sigma_0^2) - \frac{nd}{2}\log(\sigma_0^{*2}) + \frac{\|\boldsymbol{X}_1\|_F^2}{2\sigma_0^2} - \frac{\|\boldsymbol{X}_1\|_F^2}{2\sigma_0^{*2}} + \frac{\sum_{t=2}^T\|\boldsymbol{X}_t-\boldsymbol{X}_{t-1}\|_F^2}{2\tau^2} - \frac{\sum_{t=2}^T\|\boldsymbol{X}_t-\boldsymbol{X}_{t-1}\|_F^2}{2\tau^{*2}}.$$

With the constrained region $\tau^{*2} < \tau^2 < \tau^{*2}e^{\epsilon_{n,T}^2}$ and $\sigma_0^{*2} < \sigma_0^2 < \sigma_0^{*2}e^{\epsilon_{n,T}^2}$, we have,

$$\log\frac{p(\mathcal{X}|\tau^*,\sigma_0^*)}{p(\mathcal{X}|\tau,\sigma_0)} \le \frac{n(T-1)d}{2}\epsilon_{n,T}^2 + \frac{nd}{2}\epsilon_{n,T}^2 \lesssim Tn(n-1)\epsilon_{n,T}^2,$$

which implies that the third term of the KL divergence (A.17) is also bounded by $Tn(n-1)\epsilon_{n,T}^2$. Therefore, we proved that condition (A.16) is satisfied.

Finally, the conclusion holds by applying similar arguments in the final part of the proof of Theorem 7.

■

A.7 Nodewise Adaptive Priors

In this section, we consider the likelihood (1) with nodewise adaptive priors:

$$\begin{aligned} &\boldsymbol{x}_{i1} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{0i}^2\mathbb{I}_d), && \boldsymbol{x}_{i(t+1)} \mid \boldsymbol{x}_{it} \sim \mathcal{N}(\boldsymbol{x}_{it}, \tau_i^2\mathbb{I}_d), \\ &\sigma_{0i}^2 \sim \text{Inverse-Gamma}\left(a_{\sigma_0}, b_{\sigma_0}\right), && \tau_i^2 \sim \text{Gamma}\left(c_\tau, d_\tau\right), \end{aligned} \tag{A.20}$$

for $i = 1, ..., n; t = 1, ..., T-1$ to capture the nodewise level differences. The SMF are now in the following form:

$$q(\mathcal{X}, \boldsymbol{\tau}, \boldsymbol{\sigma}_0, \beta) = \prod_{i=1}^n q_i(\boldsymbol{x}_{i\cdot})q(\tau_i)q(\sigma_{0i})q(\beta). \tag{A.21}$$

First, there are only minimal changes in the computational framework. First, for the $q_i(\boldsymbol{x}_{i\cdot})$ updatings, we have the graph potentials $\boldsymbol{x}_{i\cdot}$ as follows:

$$\phi_{i1}(\boldsymbol{x}_{i1}) = \exp\{-\mu_{1/\tau_i^2}\|\boldsymbol{x}_{i1}\|_2^2/2 - \mu_{1/\sigma_{0i}^2}\|\boldsymbol{x}_{i1}\|_2^2/2\}\prod_{j\neq i}\exp[\mathbf{E}_{q(\beta)q(\boldsymbol{x}_{j1})}\{\log P_\alpha(Y_{ij1} \mid \boldsymbol{x}_{i1}, \boldsymbol{x}_{j1}, \beta)\}],$$

$$\phi_{it}(\boldsymbol{x}_{it}) = \exp\{-\mu_{1/\tau_i^2}\|\boldsymbol{x}_{it}\|_2^2/2\}\prod_{j\neq i}\exp[\mathbf{E}_{q(\beta)q(\boldsymbol{x}_{jt})}\{\log P_\alpha(Y_{ij1} \mid \boldsymbol{x}_{it}, \boldsymbol{x}_{jt}, \beta)\}],\ \forall t \in \{2, ..., T\}$$

$$\psi_{it,i(t+1)}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i(t+1)}) = \exp(\mu_{1/\tau_i^2}\boldsymbol{x}_{i(t+1)}'\boldsymbol{x}_{it}),\ \forall t \in \{1, ..., T-1\},$$

where $\mu_{1/\tau_i^2} = \mathbf{E}_{q(\tau_i)}(1/\tau_i^2)$ and $\mu_{1/\sigma_{0i}^2} = \mathbf{E}_{q(\sigma_{0i})}(1/\sigma_{0i}^2)$. Then the updating of $q_i(\boldsymbol{x}_{i\cdot})$ follows the same MP framework under the above revised potentials. In addition, for the updating of scales, we have

$$q^{(new)}(\tau_i^2) \propto \exp\left[\mathbf{E}_{q_i(\boldsymbol{x}_{i\cdot})}\left\{-\sum_{t=2}^{T}\frac{\|\boldsymbol{x}_{it}-\boldsymbol{x}_{i(t-1)}\|_2^2}{2\tau_i^2}\right\}-\frac{(T-1)d+c_\tau-1}{2}\log(\tau_i^2)-d_\tau\tau_i^2\right].$$
$$q^{(new)}(\sigma_{0i}^2) \propto \exp\left[\mathbf{E}_{q(\boldsymbol{x}_{i1})}\left(-\frac{\|\boldsymbol{x}_{i1}\|_2^2}{2\sigma_{0i}^2}\right)-\left(\frac{d}{2}+a_{\sigma_0}+1\right)\log(\sigma_{0i}^2)-\frac{b_{\sigma_0}}{\sigma_{0i}^2}\right]. \tag{A.22}$$

Therefore, we can obtain the that new update of $q(\tau_i^2)$ follows a Generalized inverse Gaussian distribution with parameter $a = 2d_\tau, b = \mathbf{E}_{q_i(\boldsymbol{x}_{i\cdot})}\{\sum_{t=2}^{T}\|\boldsymbol{x}_{it}-\boldsymbol{x}_{i(t-1)}\|_2^2/2\}, p = 1/2-(T-1)d/2-c_\tau/2$. Then the moment required in updating $\boldsymbol{x}_{it}$ can be obtained: $\mathbf{E}_{q(\tau_i)}(1/\tau_i^2) = K_{p+1}(\sqrt{b})/\left\{\sqrt{b}K_p(\sqrt{b})\right\}-2p/b$, where $K_p(\cdot)$ is the modified Bessel function of the second kind. In addition, the new update of $\sigma_{0i}^{(new)2} \sim$ Inverse-Gamma$((d+a_{\sigma_0})/2, \{\mathbf{E}_{q(\boldsymbol{x}_{i1})}(\|\boldsymbol{x}_{i1}\|_2^2)+2b_{\sigma_0}\}/2)$, which implies $\mu_{1/\sigma_{0i}^2} = \mathbf{E}_{q(\sigma_{0i})}(1/\sigma_{0i}^2) = (d+a_{\sigma_0})/\{\mathbf{E}_{q(\boldsymbol{x}_{i1})}(\|\boldsymbol{x}_{i1}\|_2^2)+2b_{\sigma_0}\}$.

To capture a smooth evolution of the latent coordinates over time for each node, we assume the following parameter space for the latent positions:

$$\text{PWAD}(\boldsymbol{L}) := \left\{\mathcal{X} : \|\boldsymbol{x}_{it}-\boldsymbol{x}_{i(t-1)}\|_2 \le \frac{L_i}{nT},\ \boldsymbol{L}=[L_1,...,L_n], L := \|\boldsymbol{L}\|_\infty\right\}, \tag{A.23}$$

where PWAD denotes point-wise adaptive dependence. The theoretical results can also be obtained similarly:

Theorem 11 (Fractional posterior convergence rate for nodewise adaptive priors) *Suppose the true data generating process satisfies Equation* (5)*, $\mathcal{X}^* \in$ PAWD($\boldsymbol{L}$) with $0 \le L = o(Tn^2)$ and condition* (8) *holds. Suppose d is a known fixed constant. Let $\epsilon_{n,T} = \|\boldsymbol{L}\|_2^{1/3}/(T^{1/3}n^{1/2}) + \sqrt{\log(nT)/(nT)}$. Then if we apply the priors defined in Equation* (2) *and adopt priors* (A.20) *for σ_{0i} and τ_i, we have for $n, T \to \infty$,*

$$\mathbf{E}\left[\Pi_\alpha\left\{\frac{1}{Tn(n-1)}\sum_{t=1}^{T}\sum_{i\neq j=1}^{n}\left(\hat{\boldsymbol{x}}_{it}'\hat{\boldsymbol{x}}_{jt}-\boldsymbol{x}_{it}^{*\prime}\boldsymbol{x}_{jt}^*\right)^2 \ge M\epsilon_{n,T}^2 \mid \mathcal{Y}\right\}\right] \to 0, \tag{A.24}$$

with $P_{\mathcal{X}^*}$ probability converging to one, where $M > 0$ is a large enough constant.

Proof The proof is similar to the proof of Theorem 4 in Section A.3. It suffices to show that the prior concentration for the set $N(\sigma_0^*, \tau^*) = \{|\sigma_{0i}^2-\sigma_0^{*2}| \le \sigma_0^{*2}/2, |\tau_i^2-\tau^{*2}| \le \tau^{*2}/2,\ i = 1,..,n\}$ is sufficiently large. Due to the independence of the prior, we have

$$-\log P(|\sigma_{0i}^2-\sigma_0^{*2}| \le \sigma_0^{*2}/2, i = 1,...,n) = -\log\{P(|\sigma_{01}^2-\sigma_0^{*2}|)^n\} \lesssim n \lesssim n(n-1)T\epsilon_{n,T}^2.$$

Similarly,

$$\begin{aligned}-\log P(|\tau_i^2-\tau^{*2}| \le \tau^{*2}/2, i = 1,...,n) &= -n\log\{P(|\tau_1^2-\tau^{*2}| \le \tau^{*2}/2)\}\\ &\lesssim -n\log(\epsilon_{n,T}L/(nT)+\log^2(nT)/(2nT^2))+n\tau^{*2}-n\log(\tau^{*2}).\end{aligned}$$

Since $\log^2(nT)/(nT^2) \le \tau^{*2} \le L/(nT)+\log^2(nT)/(nT^2)$, we have $-n\log(\log^2(nT)/(2nT^2)) \lesssim n\log(nT)$; $n\tau^{*2} \le L/T+\log^2(nT)/T^2 \lesssim n\log(nT)$ and $-n\log(\tau^{*2}) \lesssim n\log(nT)$. Therefore,

we show that $-\log \Pi(N(\sigma_0^*, \tau^*)) \lesssim n(n-1)T\epsilon_{n,T}^2$, then the rest of the proof follows the same with Section A.3 by the improved bound $\|\boldsymbol{X}_t^* - \boldsymbol{X}_{t-1}^*\|_F^2 \leq \|\boldsymbol{L}\|_2^2/(n^2T^2)$ in equation A.10. ■

Theorem 12 (Variational risk bound for nodewise adaptive SMF) *Suppose the true data generating process satisfies Equation* (5)*,* $\mathcal{X}^* \in PAWD(\boldsymbol{L})$ *with* $0 \leq L = o(Tn^2)$ *and condition* (8) *holds. Suppose* d *is a known fixed constant. Let* $\epsilon_{n,T} = \|\boldsymbol{L}\|_2^{1/3}/(T^{1/3}n^{1/2}) + \sqrt{\log(nT)/(nT)}$*. Then if we apply the priors defined in Equation* (2) *and adopt priors* (A.20) *for* σ_{oi} *and* τ_i *for* $i = 1, ..., n$ *and obtaining the optimal variational distribution* $\hat{q}(\mathcal{X})$ *under nodewise adaptive SMF family* (A.21)*, we have with* $P_{\mathcal{X}^*}$ *probability tending to one as* $n, T \to \infty$*,*

$$\mathbf{E}_{\hat{q}(\mathcal{X})}\left[\frac{1}{Tn(n-1)}\sum_{t=1}^{T}\sum_{i\neq j=1}^{n}\left(\hat{\boldsymbol{x}}_{it}'\hat{\boldsymbol{x}}_{jt} - \boldsymbol{x}_{it}^{*'}\boldsymbol{x}_{jt}^*\right)^2\right] \lesssim \epsilon_{n,T}^2. \tag{A.25}$$

Proof We consider the following variational distribution:

$$\begin{aligned} q(\mathcal{X}, \tau, \sigma_0) \propto \prod_{i=1}^{n}\prod_{t=2}^{T} p(\boldsymbol{x}_{it} \mid \boldsymbol{x}_{i(t-1)}, \tau^*)\mathbb{1}\{\|\boldsymbol{x}_{it} - \boldsymbol{x}_{it}^*\|_2 \leq c\epsilon_{n,T}\} \\ \times \prod_{i=1}^{n} p(\boldsymbol{x}_{i1} \mid \sigma_0^*)\mathbb{1}\{\|\boldsymbol{x}_{i1} - \boldsymbol{x}_{i1}^*\|_2 \leq c\epsilon_{n,T}\} \\ \times \prod_{i=1}^{n} p(\tau_i)\mathbb{1}\{\tau^{*2} < \tau_i^2 < \tau^{*2}e^{\epsilon_{n,T}^2}\}\prod_{i=1}^{n} p(\sigma_{0i})\mathbb{1}\{\sigma_0^{*2} < \sigma_{0i}^2 < \sigma_0^{*2}e^{\epsilon_{n,T}^2}\}. \end{aligned} \tag{A.26}$$

■

After the change of the priors and variational family, first by Equation (A.19) and $-n\log(\epsilon_{n,T}^2) \lesssim n\log(nT)$, we have

$$D_{KL}(q(\boldsymbol{\sigma}_0)||p(\boldsymbol{\sigma}_0)) = -n\log(P(\sigma_0^{*2} < \sigma_{01}^2 < \sigma_0^{*2}e^{\epsilon_{n,T}^2})) \lesssim -n\log(\epsilon_{n,T}^2) \lesssim Tn(n-1)\epsilon_{n,T}^2.$$

Similarly, by Equation (A.18), we also have

$$\begin{aligned} D_{KL}(q(\boldsymbol{\tau})||p(\boldsymbol{\tau})) &= -n\log(P(\tau^{*2} < \tau_1^2 < \tau^{*2}e^{\epsilon_{n,T}^2})) \\ &\lesssim -n\log(\epsilon_{n,T}^2) - n\log(\min_{\tau^{*2}<\tau^2<\tau^{*2}e^{\epsilon_{n,T}^2}} f_{c_\tau,d_\tau}(\tau^2)) \\ &\lesssim Tn(n-1)\epsilon_{n,T}^2. \end{aligned}$$

Moreover, we have the density,

$$\begin{aligned} p(\mathcal{X} \mid \boldsymbol{\tau}, \boldsymbol{\sigma}_0) = \frac{1}{(\sqrt{2\pi})^{nTd}}\exp\Bigg\{-\sum_{i=1}^{n}\frac{(T-1)d}{2}\log(\tau_i^2) - \sum_{i=1}^{n}\frac{d}{2}\log(\sigma_{0i}^2) \\ -\sum_{i=1}^{n}\frac{\|\boldsymbol{x}_{i1}\|_2^2}{2\sigma_{0i}^2} - \frac{\sum_{i=1}^{n}\sum_{t=2}^{T}\|\boldsymbol{x}_{it} - \boldsymbol{x}_{i(t-1)}\|_2^2}{2\tau_i^2}\Bigg\}, \end{aligned}$$

which implies that

$$\log\frac{p(\mathcal{X}|\boldsymbol{\tau}^*,\boldsymbol{\sigma}_0^*)}{p(\mathcal{X}|\boldsymbol{\tau},\boldsymbol{\sigma}_0)}=\sum_{i=1}^{n}\frac{(T-1)d}{2}\log(\tau_i^2)-\frac{n(T-1)d}{2}\log(\tau^{*2})+\sum_{i=1}^{n}\frac{d}{2}\log(\sigma_{0i}^2)-\frac{nd}{2}\log(\sigma_0^{*2})$$
$$+\sum_{i=1}^{n}\frac{\|\boldsymbol{x}_{i1}\|_2^2}{2\sigma_{0i}^2}-\frac{\|\boldsymbol{X}_1\|_F^2}{2\sigma_0^{*2}}+\frac{\sum_{i=1}^{n}\sum_{t=2}^{T}\|\boldsymbol{x}_{it}-\boldsymbol{x}_{i(t-1)}\|_2^2}{2\tau_i^2}-\frac{\sum_{t=2}^{T}\|\boldsymbol{X}_t-\boldsymbol{X}_{t-1}\|_F^2}{2\tau^{*2}},$$

where $\boldsymbol{\tau}^*=(\tau^*,\tau^*,...,\tau^*)'$ and $\boldsymbol{\sigma}_0^*=(\sigma_0^*,\sigma_0^*,...,\sigma_0^*)'$. With the constrained region $\tau^{*2}<\tau^2<\tau^{*2}e^{\epsilon_{n,T}^2}$ and $\sigma_0^{*2}<\sigma_0^2<\sigma_0^{*2}e^{\epsilon_{n,T}^2}$, we have,

$$\log\frac{p(\mathcal{X}|\boldsymbol{\tau}^*,\boldsymbol{\sigma}_0^*)}{p(\mathcal{X}|\boldsymbol{\tau},\boldsymbol{\sigma}_0)}\leq\frac{n(T-1)d}{2}\epsilon_{n,T}^2+\frac{nd}{2}\epsilon_{n,T}^2\lesssim Tn(n-1)\epsilon_{n,T}^2.$$

Then the rest of the proofs follow the same with proof of Theorem (7) in Section A.6.

A.8 Auxiliary Lemmas

Lemma 13 (Small ball probability, Theorem 1.1 in Shao, 1993) *Let $\{X(t),0\leq t\leq 1\}$ be a real-valued Gaussian process with mean zero, $X(0)=0$ and stationary increments. Denote $\sigma^2(h)=E(X(t+h)-X(t))^2$ for $0\leq t\leq t+h\leq 1$. If $\sigma^2(h)$ is concave and $\sigma(h)/h^\alpha$ is non-decreasing in $(0,1)$ for some $\alpha>0$, then we have*

$$P(\sup_{0\leq t\leq 1}|X(t)|\leq C_\alpha\sigma(x))\geq\exp(-2/x),$$

where $C_\alpha=1+3e\sqrt{\pi/\alpha}$.

Lemma 14 (Upper bound for binary KL divergence) *Let $p_a=1/(1+\exp(-a))$ and $p_b=1/(1+\exp(-b))$. Define P_a and P_b as the Bernoulli measures with probability p_a and p_b. Then we have*

$$D_{KL}(P_a\,||\,P_b)+D_{KL}(P_b\,||\,P_a)\leq(p_a\vee p_b)(a-b)^2.$$

Proof

$$D_{KL}(P_a\,||\,P_b)+D_{KL}(P_b\,||\,P_a)=(p_a-p_b)\log\frac{p_a}{p_b}+(p_b-p_a)\log\frac{1-p_a}{1-p_b}$$
$$=(p_a-p_b)\log\left(\frac{p_a}{1-p_a}\frac{1-p_b}{p_b}\right)=\left\{\frac{1}{1+\exp(-a)}-\frac{1}{1+\exp(-b)}\right\}\log(e^ae^{-b})$$
$$=(a-b)\left\{\frac{1}{1+\exp(-a)}-\frac{1}{1+\exp(-b)}\right\}.$$

Without loss of generality, we can assume $a>b$, then by $\exp(x)\geq 1+x$, we have

$$\frac{1}{1+\exp(-a)}-\frac{1}{1+\exp(-b)}=\frac{e^{-b}-e^{-a}}{(1+\exp(-a))(1+\exp(-b))}$$
$$\leq\frac{1-e^{b-a}}{(1+e^{-a})(1+e^b)}\leq p_a(1-e^{b-a})\leq p_a(a-b).$$

■

Lemma 15 (Upper bound of second order KL moment) *Let $p_a = 1/(1+\exp(-a))$ and $p_b = 1/(1+\exp(-b))$. Define P_a and P_b as the Bernoulli measures with probability p_a and p_b. Then we have*

$$\int P_a \log^2\left(\frac{P_a}{P_b}\right) d\mu \le \left[\frac{p_a}{(p_a \wedge p_b)^2} + \frac{1-p_a}{(1-p_a \vee p_b)^2}\right](p_a \vee p_b)^2(a-b)^2.$$

Proof Note that

$$\int P_a \log^2\left(\frac{P_a}{P_b}\right) d\mu = p_a \log^2\left(\frac{p_a}{p_b}\right) + (1-p_a)\log^2\left(\frac{1-p_a}{1-p_b}\right).$$

We have

$$\log^2\left(\frac{p_a}{p_b}\right) = \log^2\left(\frac{p_a \vee p_b}{p_a \wedge p_b} - 1 + 1\right) \le \left(\frac{p_a \vee p_b - p_a \wedge p_b}{p_a \wedge p_b}\right)^2 = \left(\frac{p_a - p_b}{p_a \wedge p_b}\right)^2.$$

Similarly,

$$\log^2\left(\frac{1-p_a}{1-p_b}\right) = \log^2\left(\frac{(1-p_a)\vee(1-p_b)}{(1-p_a)\wedge(1-p_b)} - 1 + 1\right) \le \left(\frac{p_a - p_b}{1-p_a \vee p_b}\right)^2.$$

For the $(p_a - p_b)^2$ term, by $\exp(x) \ge 1 + x$, we have

$$\begin{aligned}
&\frac{1}{1+\exp(-a\vee b)} - \frac{1}{1+\exp(-a\wedge b)} = \frac{e^{-a\wedge b} - e^{-a\vee b}}{(1+\exp(-a\vee b))(1+\exp(-a\wedge b))} \\
\le\ &\frac{1-e^{a\vee b - a\wedge b}}{(1+e^{a\wedge b})(1+e^{-a\vee b})} \le (p_a \vee p_b)(1-e^{a\vee b - a\wedge b}) \le (p_a \vee p_b)(a\wedge b - a\vee b).
\end{aligned}$$

■

Lemma 16 (Lower bound of the $1/2$ divergence) *Let $p_a = 1/(1+\exp(-a))$ and $p_b = 1/(1+\exp(-b))$. Define P_a and P_b as the Bernoulli measures with probability p_a and p_b.*

1. *Suppose that there exist constants $c, C > 0$ such that $c < a, b < C$, then we have*

$$D_{\frac{1}{2}}(P_a, P_b) \gtrsim (b-a)^2.$$

2. *Suppose that $a, b \to -\infty$ such that $p_a, p_b \to 0$, then we have*

$$D_{\frac{1}{2}}(P_a, P_b) \gtrsim \exp\{a \wedge b\}(b-a)^2.$$

Proof

$$D_{\frac{1}{2}}(p_a, p_b) = -2\log(1 - h^2(p_a, p_b)) \ge 2h^2(p_a, p_b) = \left[(\sqrt{p_a} - \sqrt{p_b})^2 + (\sqrt{1-p_a} - \sqrt{1-p_b})^2\right].$$

For the first conclusion, since a, b are bounded, p_a, p_b are bounded away form 0 and 1, and $(\sqrt{p_a}+\sqrt{p_b}), (\sqrt{1-p_a}+\sqrt{1-p_b})$ are bounded from 0 as well. Hence,

$$D_{\frac{1}{2}}(p_a, p_b) \gtrsim \left[(\sqrt{p_a}-\sqrt{p_b})^2(\sqrt{p_a}+\sqrt{p_b})^2 + (\sqrt{1-p_a}-\sqrt{1-p_b})^2(\sqrt{1-p_a}+\sqrt{1-p_b})^2\right]$$
$$\gtrsim (p_a - p_b)^2 \overset{(i)}{=} \left\{\frac{\exp(x)}{(1+\exp(x))^2}\right\}^2 (a-b)^2 \gtrsim (a-b)^2.$$

where (i) is because the mean value theorem and $a < x < b$ is bounded.

For the second conclusion, when the probabilities p_a, p_b are converging to zeros, for the term $(\sqrt{p_a}-\sqrt{p_b})^2$, by the mean value theorem of function $\sqrt{p_x}$ with respect to x, we have

$$(\sqrt{p_a}-\sqrt{p_b})^2 \geq \left(\frac{\sqrt{\exp(x)}\sqrt{1+\exp(x)}}{2(1+\exp(x))^2}\right)^2 (a-b)^2 \overset{(i)}{\gtrsim} e^{a\wedge b}(a-b)^2,$$

where (i) is because $\exp(x)$ is the order of $\exp\{a \wedge b\}$ for $a \wedge b < x < a \vee b$. For the term $(\sqrt{1-p_a}-\sqrt{1-p_b})^2$, note that $\sqrt{1-p_a}+\sqrt{1-p_b}$ is still bound away from 0, we have

$$(\sqrt{1-p_a}-\sqrt{1-p_b})^2 \gtrsim (p_a-p_b)^2 = \left\{\frac{\exp(x)}{(1+\exp(x))^2}\right\}^2 (a-b)^2 \gtrsim e^{(2a)\wedge(2b)}(a-b)^2,$$

for $a < x < b$. Finally, $\exp(a \wedge b)(a-b)^2$ dominates when the sum of the two lower bounds is taken into account.

■

Lemma 17 (Probability bound for maximal of sub-Gaussian random variables) *Let $X_1, ..., X_n$ be independent sub-Gaussian random variables with mean zero and sub-Gaussian norm upper bounded by σ. Then we have for every $t > 0$,*

$$P\left\{\max_{i=1,...,n} |X_i| \geq \sqrt{2\sigma^2(\log n + t)}\right\} \leq 2e^{-t}.$$

Proof By union bound and the sub-Gaussianity, we have

$$P\left\{\max_{i=1,...,n} |X_i| \geq u\right\} \leq \sum_{i=1}^{n} P\{|X_i| \geq u\} \leq 2ne^{-\frac{u^2}{2\sigma^2}},$$

by choosing $u = \sqrt{2\sigma^2(\log n + t)}$, the conclusion is proved. ■

A.9 Additional Simulation Examples

Gaussian Networks: 25 replicated data sets are generated from $Y_{ijt} \sim \mathcal{N}(0.1+\boldsymbol{x}_{it}'\boldsymbol{x}_{jt}, 0.1^2)$ for $i \neq j = 1, ..., n$ and $t = 1, ..., T$ where $n = 100$, $T = 100$, $d = 2$. Let $\boldsymbol{x}_{i1} \sim \mathcal{N}((0,0)', \tau^2\mathbb{I})$, and transitions $\boldsymbol{x}_{it} = \boldsymbol{x}_{i(t-1)} + \boldsymbol{\epsilon}_{t-1}$, where given any coordinate j for a fixed node i, let $[\epsilon_{ij1}, ..., \epsilon_{ijT}]' \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbb{I})$. The iterations are stopped when the difference between predictive RMSEs in two consecutive cycles is less than 10^{-3}. In the algorithm, both the initial

τ	0.001	0.005	0.01	0.05	0.1	0.5
MF	0.0101	0.0105	0.0129	0.0219	0.0205	0.0211
SMF	2.34×10^{-3}	5.97×10^{-3}	9.53×10^{-3}	0.0200	0.0206	0.0210

Table 2: Performance comparison for Gaussian networks between SMF and MF. The measure is the median of root mean square error for estimation of the latent distances of the repeated simulations.

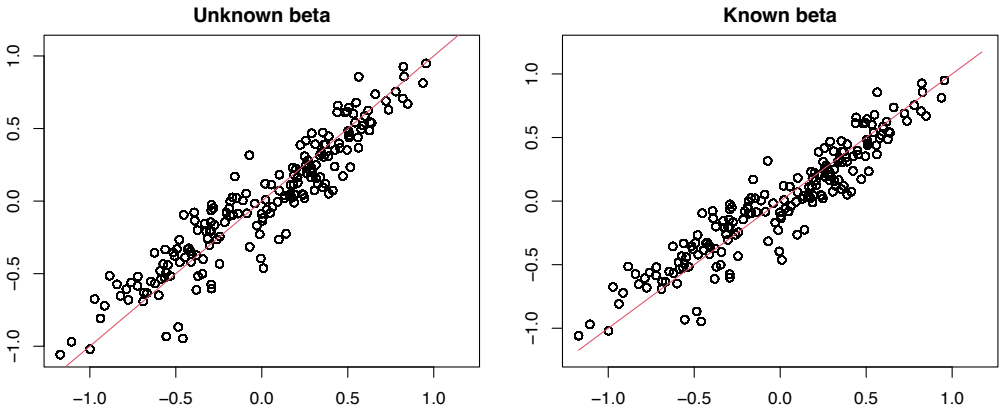

Figure 9: Comparison of recovery of the inner product $\mathbf{x}_{it}^{*\prime}\mathbf{x}_{it}^{*}$ for estimating $\hat{\beta}$ using the algorithm in the left vs correctly specifying $\hat{\beta} = 0$ in the right. The algorithm estimates $\hat{\beta} = 0.0148$ when β is assumed unknown. The x-axis is the true inner products $\boldsymbol{x}_{it}^{*'}\boldsymbol{x}_{jt}^{*}$ and the y-axis is the estimated inner products $\hat{\boldsymbol{x}}_{it}'\hat{\boldsymbol{x}}_{jt}$.

and transition variances are learned adaptively with prior (3). The prior for the intercept is set to $\mathcal{N}(0, 10)$. Table 2 shows the mean of the 25 replicated simulations. Clearly, SMF performs much better than MF in parameter recovery when the transition is small. The result again reinforces that when the dependence among latent positions is significant, SMF should be adopted.

Reused simulation case: For the many simulation cases in this subsection, we use 25 replicated data sets generated from the following case: $Y_{ijt} \sim \mathcal{N}(\boldsymbol{x}_{it}'\boldsymbol{x}_{jt}, 0.1^2)$ for $i \neq j = 1, ..., n$ and $t = 1, ..., T$ where $n = 20$, $T = 20$, $d = 2$. Let $\boldsymbol{x}_{i1} \sim \mathcal{N}((0, 0)', \tau^2\mathbb{I})$, and transitions $\boldsymbol{x}_{it} = \boldsymbol{x}_{i(t-1)} + \boldsymbol{\epsilon}_{t-1}$, where given any coordinate j for a fixed node i, let $[\epsilon_{ij1}, ..., \epsilon_{ijT}]' \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbb{I})$. The prior is set in the same way as the above.

Recovery of inner products: We consider a simulation case regarding visualization of the recovery of $\boldsymbol{x}_{it}'\boldsymbol{x}_{jt}$, by showing the comparison of estimated $\hat{\boldsymbol{x}}_{it}'\hat{\boldsymbol{x}}_{jt}$ vs the truth $\boldsymbol{x}_{it}^{*'}\boldsymbol{x}_{jt}^{*}$. The data is generated for one realization of the 'Reused simulation case' with $\beta^* = 0$ and the two following scenarios: 1. Using our algorithm with estimating $\hat{\beta}$ as unknown beta case; 2. Specify $\hat{\beta} = 0$ as known beta case. We then compare the true inner products $\boldsymbol{x}_{it}^{*'}\boldsymbol{x}_{jt}^{*}$ and the estimated inner products $\hat{\boldsymbol{x}}_{it}'\hat{\boldsymbol{x}}_{jt}$ without adding the estimated intercept for both cases. The simulation is provided in Figure 9: In the figure above, it is evident that the

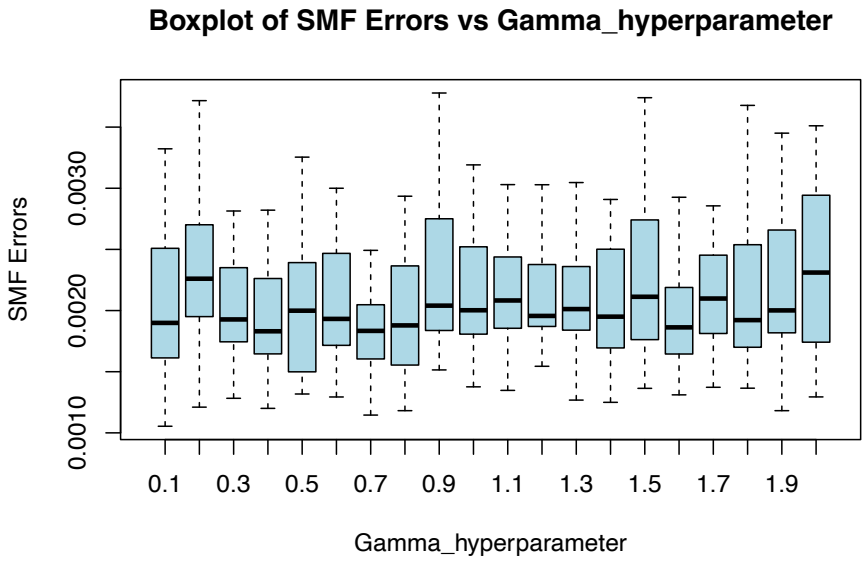

Figure 10: Performance comparison for Gaussian networks of estimation of SMF and MF across different hyperparameters of d_τ for Gamma distribution with $c_\tau = 1$.

estimation of inner products remains accurate even when the value of β is estimated in the algorithm instead of being correctly specified at $\beta^* = 0$. This is due to the good estimate of $\hat{\beta} = 0.0148$. The theoretical explanations behind this intriguing phenomenon are reserved for future research.

Sensitivity analysis with respect to the choice of Hyperparameters: For the Gamma prior, we test the sensitivity of the hyperparameters c_τ and d_τ. We fix $c_\tau = 1$ and change d_τ from 0.1 to 2 with 20 grids and repeat 25 simulations of the 'Reused simulation case' and the result is shown in Figure 10. On the other hand, we also fix $d_\tau = 1/2$ and change c_τ from 0.1 to 2 with 20 grids and repeat 25 simulations of the 'Reused simulation case', and the result is shown in Figure 11.

Sensitivity analysis with respect to the choice of α: We compared the effect of different α values in our model. We use 11 grids for α
$=0.5,\ 0.55,\ 0.6,\ 0.65,\ 0.7,\ 0.75,\ 0.8,\ 0.85,\ 0.9,\ 0.95, 0.99$. We repeat the simulation of the 'Reused simulation case' for 25 times. The simulation result is shown in Figure 12. As can be seen from the figure, the results are consistent and suggest that the choice of α does not affect the outcome.

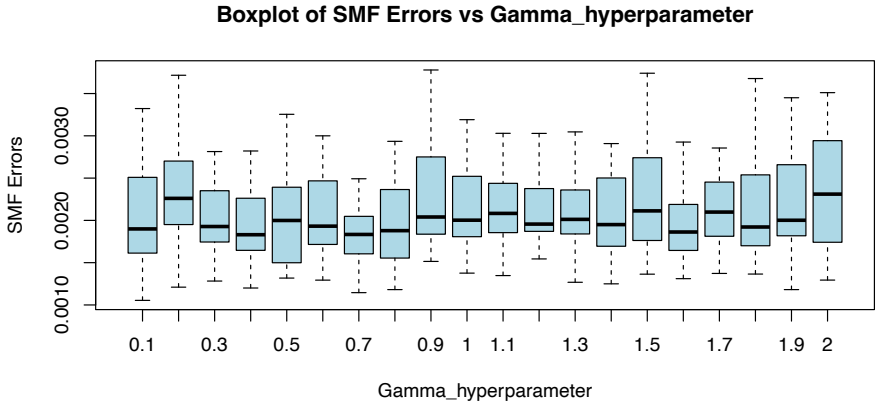

Figure 11: Performance comparison for Gaussian networks of estimation of SMF and MF across different hyperparameters of c_τ for Gamma distribution with $d_\tau = 1/2$.

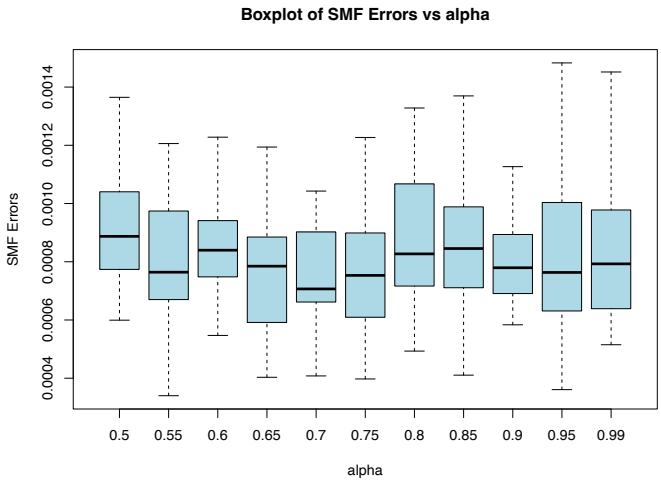

Figure 12: Performance comparison for Gaussian networks of estimation of SMF and MF across different α.

Sensitivity analysis with respect to the choice of latent dimensions for Enron's email data: In the Enron email data set, we found that the $d = 5$ case provides a good comparison between our method and using an inverse Gamma prior. Our method consistently performed better than using the inverse Gamma prior. Here we have also shown the comparison results for $d = 2, 3, 4$ in Figure 13. We observed that $d = 2, 3, 4, 5$ showed similar behavior, where using the Gamma prior consistently outperformed the inverse Gamma prior for the transition variance.

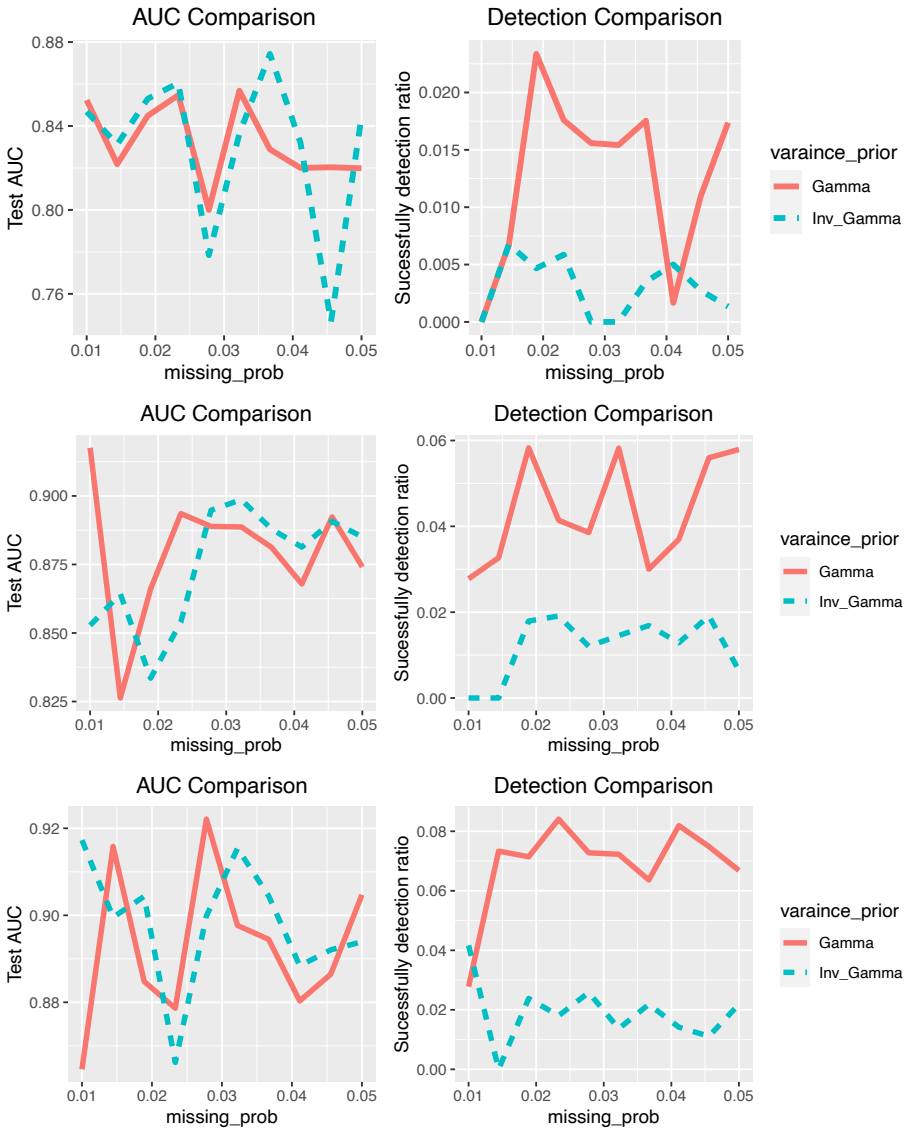

Figure 13: Comparison using Enron's email data set between Gamma prior and Inverse Gamma prior on the transition variance, with cases for $d = 2, 3, 4$ displayed from top to bottom.

Nodewise adaptive priors: Our new simulations demonstrate that node adaptivity can improve estimation accuracy when different nodes have different transition scales. We consider a scenario where 90% nodes remain static across all time (so that their corresponding L_i can be seen as 0), while only the rest 10% of nodes change over time. For the changing node, we use τ as the true transition standard derivation to control the magnitude of changes: $\boldsymbol{x}_{it} \mid \boldsymbol{x}_{i(t-1)} \sim \mathcal{N}(\boldsymbol{x}_{i(t-1)}, \tau^2 \boldsymbol{I}_d)$. The other settings are similar: we use 25 replicated data sets are generated from $Y_{ijt} \sim \text{Bernoulli}(\boldsymbol{x}'_{it}\boldsymbol{x}_{jt})$ for $i \neq j = 1, ..., n$ and

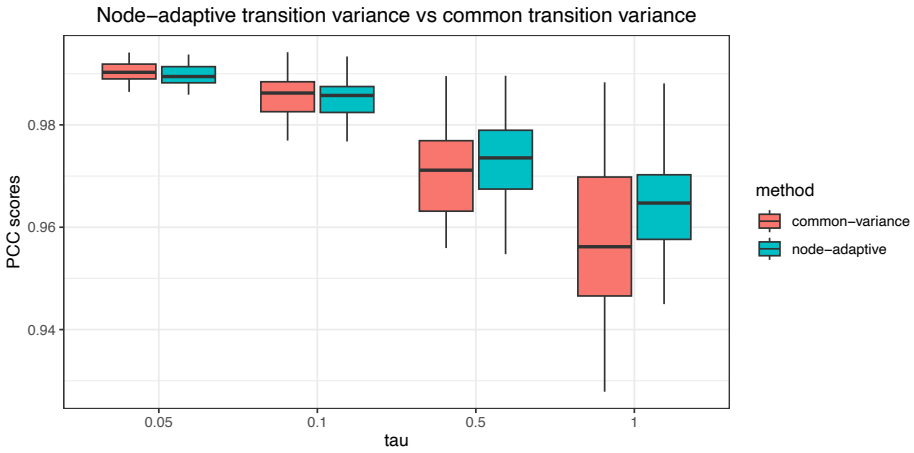

Figure 14: Performance comparison for nodewise adaptive prior on variance vs. common transition variance across different τ as the transition standard derivation 10% of changing nodes, while the rest 90% of nodes stay static across all time points.

$t = 1, ..., T$ where $n = 20, T = 50, d = 2$. Let $\boldsymbol{x}_{i1} \sim \mathcal{N}((0,0)', 0.1\mathbb{I})$, where given any coordinate j for a fixed node i. We compare the estimation accuracy between the nodewise adaptive priors (4) and common variance priors (3) across different values of τ such as $\tau = 0.05, 0.1, 0.5, 1$. Figure 14 illustrates the performance between nodewise adaptive priors and common priors. When $\tau = 0.05, 0.1$ is small, the common variance prior performs most the same or slightly better than nodewise adaptive priors because differences in the magnitude of change for different nodes may not be large enough and both methods may converge at the same rate. However, when $\tau = 0.5, 1$ are not close to zero, nodewise adaptive priors perform much better than the common variance prior. This is reasonable as the differences in the change of scale between different nodes are large in these cases and introducing nodewise adaptivity can capture the true data-generating process more precisely.

A.10 MF Updatings for β

Suppose the prior for β is $\mathcal{N}(\mu_\beta, \sigma_\beta^2)$. For Gaussian likelihood, the updating for β can be obtained

$$\begin{aligned}
\hat{q}(\beta) &\propto \exp[\mathbf{E}_{-\beta}\{\log p_\alpha(\mathcal{X}, \beta, \mathcal{Y})\}] \propto \exp[\mathbf{E}_{-\beta}\{\log P_\alpha(\mathcal{Y} \mid \mathcal{X}, \beta)\} + \log p(\beta)] \\
&\propto \exp\left[\mathbf{E}_{-\beta}\left\{\alpha \sum_{t=1}^{T} \sum_{i \neq j} -\frac{(Y_{ijt} - \beta - \boldsymbol{x}_{it}' \boldsymbol{x}_{jt})^2}{2\sigma^2}\right\} - \frac{(\beta - \mu_\beta)^2}{2\sigma_\beta^2}\right] \\
&\propto \exp\left[\sum_{t=1}^{T} \sum_{i \neq j} -\alpha \frac{\beta^2 - 2\beta(Y_{ijt} - \boldsymbol{\mu}_{it}' \boldsymbol{\mu}_{jt})}{2\sigma^2} - \frac{(\beta - \mu_\beta)^2}{2\sigma_\beta^2}\right].
\end{aligned}$$

Therefore, $q^{(\text{new})}(\beta)$ is the density of $\mathcal{N}(\mu_\beta^{(new)}, \sigma_\beta^{(new)})$, with

$$\sigma_\beta^{(new)2} = \left\{\sigma_\beta^{-2} + \alpha T n(n-1)\sigma^{-2}\right\}^{-1}, \quad \mu_\beta^{(new)} = \sigma_\beta^{(new)^2}\left\{\sigma_\beta^{-2}\mu_\beta + \sum_{i\neq j}\sum_t \alpha\sigma^{-2}(Y_{ijt} - \boldsymbol{\mu}_{it}'\boldsymbol{\mu}_{jt})\right\}.$$

For the binary case, the updating for β after tangent transformation can also be obtained

$$q^{(\text{new})}(\beta;\Xi) \propto \exp[\mathbf{E}_{\mathcal{X}}\{\log \underline{P}_\alpha(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\beta};\Xi)\} + \log p(\beta)]$$
$$\propto \exp\left[\sum_{i\neq j}\sum_t \alpha\left\{A(\xi_{ijt})\right\}\beta^2 + \sum_{i\neq j}\sum_t \alpha\left\{Y_{ijt} - \frac{1}{2} + 2A(\xi_{ijt})\boldsymbol{\mu}_{it}'\boldsymbol{\mu}_{jt}\right\}\beta - \frac{1}{2}\sigma_\beta^{-2}\beta^2 + \mu_\beta\sigma_\beta^{-2}\beta\right].$$

Therefore, $q^{(\text{new})}(\beta;\Xi)$ is the density of $\mathcal{N}(\mu_\beta^{(new)}, \sigma_\beta^{(new)})$, with

$$\sigma_\beta^{(new)2} = \left\{\sigma_\beta^{-2} - 2\alpha\sum_{i\neq j}\sum_t A(\xi_{ijt})\right\}^{-1}, \quad \mu_\beta^{(new)} = \sigma_\beta^{(new)^2}\left[\sigma_\beta^{-2}\mu_\beta + \sum_{i\neq j}\sum_t \alpha\left\{Y_{ijt} - \frac{1}{2} + 2A(\xi_{ijt})\boldsymbol{\mu}_{it}'\boldsymbol{\mu}_{jt}\right\}\right].$$

A.11 MF Updatings for $\mathcal{X}$

The updating for β in MF is the same with SMF. For updating $\mathcal{X}$ in the Gaussian case, we have

$$\hat{q}(\boldsymbol{x}_{it}) \propto \exp[\mathbf{E}_{-\boldsymbol{x}_{it}}\{\log p_\alpha(\mathcal{X}, \beta, \mathcal{Y})] \propto \exp[\mathbf{E}_{-\boldsymbol{x}_{it}}\{\log P_\alpha(\mathcal{Y} \mid \mathcal{X}, \beta)\} + \log p(\mathcal{X})\}]$$
$$\propto \exp\left[\mathbf{E}_{-\boldsymbol{x}_{it}}\left\{\sum_{i\neq j} -\alpha\frac{(Y_{ijt} - \beta - \boldsymbol{x}_{it}'\boldsymbol{x}_{jt})^2}{2\sigma^2} - \frac{\|\boldsymbol{x}_{it} - \boldsymbol{x}_{i(t-1)}\|^2}{2\tau^2} - \frac{\|\boldsymbol{x}_{it} - \boldsymbol{x}_{i(t+1)}\|^2}{2\tau^2}\right\}\right]$$
$$\propto \exp\left[\left\{\sum_{i\neq j} -\alpha\frac{-2(Y_{ijt} - \mu_\beta^{(new)})\boldsymbol{x}_{it}'\boldsymbol{\mu}_{jt} + \boldsymbol{x}_{it}'(\boldsymbol{\mu}_{jt}\boldsymbol{\mu}_{jt}' + \boldsymbol{\Sigma}_{jt})\boldsymbol{x}_{it}}{2\sigma^2} - \frac{\|\boldsymbol{x}_{it}\|^2 - 2\boldsymbol{x}_{it}\boldsymbol{\mu}_{i(t-1)}}{2\tau^2}\right.\right.$$
$$\left.\left. - \frac{\|\boldsymbol{x}_{it}\|^2 - 2\boldsymbol{x}_{it}\boldsymbol{\mu}_{i(t+1)}}{2\tau^2}\right\}\right].$$

Therefore, $q^{(\text{new})}(\boldsymbol{x}_{it})$ is the density of $\mathcal{N}(\boldsymbol{\mu}_{it}^{(new)}, \boldsymbol{\Sigma}_{it}^{(new)})$, with

$$\boldsymbol{\Sigma}_{it}^{(new)} = \left\{2\tau^{-2}\mathbb{I} + \alpha\sigma^{-2}\sum_{i\neq j}(\boldsymbol{\mu}_{jt}\boldsymbol{\mu}_{jt}' + \boldsymbol{\Sigma}_{jt})\right\}^{-1},$$
$$\boldsymbol{\mu}_{it}^{(new)} = \boldsymbol{\Sigma}_{it}^{(new)}\left(\tau^{-2}\boldsymbol{\mu}_{i(t-1)} + \tau^{-2}\boldsymbol{\mu}_{i(t+1)} + \sum_{i\neq j}\alpha\sigma^{-2}(Y_{ijt} - \mu_\beta^{(new)})\boldsymbol{\mu}_{jt}\right).$$

For the binary case, here we derive the updating formula under the mean-filed approximation for $\mathcal{X}$ after performing the tangent approximation. For the mean-field updating for

$\boldsymbol{x}_{it}$, we have:

$$\begin{aligned}
\hat{q}(\boldsymbol{x}_{it}) &\propto \exp[\mathbf{E}_{-\boldsymbol{x}_{it}}\{\log \underline{p}_{\alpha}(\mathcal{X}, \beta, \mathcal{Y})] \propto \exp[\mathbf{E}_{-\boldsymbol{x}_{it}}\{\log \underline{P}_{\alpha}(\mathcal{Y} \mid \mathcal{X}, \beta)\} + \log p(\mathcal{X})\}] \\
&\propto \exp\left[\mathbf{E}_{-\boldsymbol{x}_{it}}\left\{\sum_{i\neq j}\alpha\left(A(\xi_{ijt})(\boldsymbol{x}_{it}'\boldsymbol{x}_{jt})^2 + (2A(\xi_{ijt})\beta + Y_{ijt} - \frac{1}{2})\boldsymbol{x}_{it}'\boldsymbol{x}_{jt}\right)\right.\right. \\
&\left.\left. - \frac{\|\boldsymbol{x}_{it} - \boldsymbol{x}_{i(t-1)}\|^2}{2\tau^2} - \frac{\|\boldsymbol{x}_{it} - \boldsymbol{x}_{i(t+1)}\|^2}{2\tau^2}\right\}\right] \\
&\propto \exp\left[\left\{\sum_{i\neq j}\alpha\left(A(\xi_{ijt})\boldsymbol{x}_{it}'(\boldsymbol{\mu}_{jt}\boldsymbol{\mu}_{jt}' + \boldsymbol{\Sigma}_{jt})\boldsymbol{x}_{it} + (2A(\xi_{ijt})\mu_{\beta}^{(new)} + Y_{ijt} - \frac{1}{2})\boldsymbol{x}_{it}'\boldsymbol{\mu}_{jt}\right)\right.\right. \\
&\left.\left. - \frac{\|\boldsymbol{x}_{it}\|^2 - 2\boldsymbol{x}_{it}\boldsymbol{\mu}_{i(t-1)}}{2\tau^2} - \frac{\|\boldsymbol{x}_{it}\|^2 - 2\boldsymbol{x}_{it}\boldsymbol{\mu}_{i(t+1)}}{2\tau^2}\right\}\right].
\end{aligned}$$

Therefore, $q^{(\text{new})}(\boldsymbol{x}_{it})$ is the density of $\mathcal{N}(\boldsymbol{\mu}_{it}^{(new)}, \boldsymbol{\Sigma}_{it}^{(new)})$, with

$$\boldsymbol{\Sigma}_{it}^{(new)} = \left\{2\tau^{-2}\mathbb{I} - 2\alpha\sum_{i\neq j}\left(A(\xi_{ijt})(\boldsymbol{\mu}_{jt}\boldsymbol{\mu}_{jt}' + \boldsymbol{\Sigma}_{jt})\right)\right\}^{-1},$$

$$\boldsymbol{\mu}_{it}^{(new)} = \boldsymbol{\Sigma}_{it}^{(new)}\left(\tau^{-2}\boldsymbol{\mu}_{i(t-1)} + \tau^{-2}\boldsymbol{\mu}_{i(t+1)} + \sum_{i\neq j}\alpha(2A(\xi_{ijt})\mu_{\beta}^{(new)} + Y_{ijt} - \frac{1}{2})\boldsymbol{\mu}_{jt}\right).$$

References

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 670–688. IEEE, 2015.

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475–1497, 2020.

Jincheng Bai, Qifan Song, and Guang Cheng. Nearly optimal variational inference for high dimensional regression with shrinkage priors. *arXiv preprint arXiv:2010.12887*, 2020.

David Barber and Silvia Chiappa. Unified inference for variational bayesian linear Gaussian state-space models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

Skye Bender-deMoll and Martina Morris. *ndtv: Network Dynamic Temporal Visualizations*, 2021. URL `https://CRAN.R-project.org/package=ndtv`. R package version 0.13.1.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Bayesian fractional posteriors. *The Annals of Statistics*, 47(1):39–66, 2019.

Aharon Birnbaum, Iain M Johnstone, Boaz Nadler, and Debashis Paul. Minimax bounds for sparse PCA with noisy high-dimensional data. *The Annals of statistics*, 41(3):1055, 2013.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Carter T. Butts, Ayn Leslie-Cook, Pavel N. Krivitsky, and Skye Bender-deMoll. *networkDynamic: Dynamic Extensions for Network Objects*, 2020. URL `https://CRAN.R-project.org/package=networkDynamic`. R package version 0.10.1.

David L Donoho and Iain M Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921, 1998.

Daniele Durante and David B Dunson. Bayesian inference and testing of group differences in brain networks. *Bayesian Analysis*, 13(1):29–58, 2018.

Daniele Durante and Tommaso Rigon. Conditionally conjugate mean-field variational Bayes for logistic models. *Statistical Science*, 34(3):472–485, 2019.

Daniele Durante, David B Dunson, and Joshua T Vogelstein. Nonparametric Bayes modeling of populations of networks. *Journal of the American Statistical Association*, 112(520): 1516–1530, 2017a.

Daniele Durante, Nabanita Mukherjee, and Rebecca C Steorts. Bayesian learning of dynamic multilayer networks. *The Journal of Machine Learning Research*, 18(1):1414–1442, 2017b.

Nial Friel, Riccardo Rastelli, Jason Wyse, and Adrian E Raftery. Interlocking directorates in Irish companies using a latent space model for bipartite networks. *Proceedings of the National Academy of Sciences*, 113(24):6629–6634, 2016.

Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

Chao Gao, Yu Lu, Zongming Ma, and Harrison H Zhou. Optimal estimation and completion of matrices with biclustering structures. *The Journal of Machine Learning Research*, 17 (1):5602–5630, 2016.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534, 2006.

Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, pages 500–531, 2000.

Indrajit Ghosh, Anirban Bhattacharya, and Debdeep Pati. Statistical optimality and stability of tangent transform algorithms in logit models. *The Journal of Machine Learning Research*, 23(184):1–42, 2022.

Manuel Gil, Fady Alajaji, and Tamas Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131, 2013.

Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, Edoardo M Airoldi, et al. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2): 129–233, 2010.

Paul Gustafson, Shahadut Hossain, and Ying C Macnab. Conservative prior distributions for variance parameters in hierarchical models. *Canadian Journal of Statistics*, 34(3): 377–390, 2006.

Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

Peter Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, pages 657–664. Curran Associates, Inc., 2008.

Peter D Hoff. Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics*, 9(3):1169, 2015.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

Tommi S Jaakkola and Michael I Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

Seonghyun Jeong and Subhashis Ghosal. Posterior contraction in sparse generalized linear models. *Biometrika*, 108(2):367–379, 2021.

Bent Jorgensen. *Statistical Properties of the Generalized Inverse Gaussian Distribution*, volume 9. Springer Science & Business Media, 2012.

R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960. ISSN 0021-9223. doi: 10.1115/1.3662552. URL `https://doi.org/10.1115/1.3662552`.

Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pages 217–226. Springer, 2004.

Olga Klopp, Alexandre B Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354, 2017.

Pavel N Krivitsky, Mark S Handcock, Adrian E Raftery, and Peter D Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31(3):204–213, 2009.

Antonio R Linero and Yun Yang. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(5):1087–1110, 2018.

Yan Liu and Yuguo Chen. Variational inference for latent space models for dynamic networks. *Statistica Sinica*, 32(4):2147–2170, 2022.

Joshua Daniel Loyal. Fast variational inference of latent space models for dynamic networks using Bayesian P-splines. *arXiv preprint arXiv:2401.09715*, 2024.

Joshua Daniel Loyal and Yuguo Chen. An eigenmodel for dynamic multilayer networks. *The Journal of Machine Learning Research*, 24(128):1–69, 2023.

Zhuang Ma, Zongming Ma, and Hongsong Yuan. Universal latent space model fitting for large networks with edge covariates. *The Journal of Machine Learning Research*, 21(4): 1–67, 2020.

Martin Maechler. *Bessel: Computations and Approximations for Bessel Functions*, 2019. URL `https://CRAN.R-project.org/package=Bessel`. R package version 0.6-0.

Enno Mammen and Sara van de Geer. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997.

Ryan Martin and Yiqi Tang. Empirical priors for prediction in sparse high-dimensional linear regression. *The Journal of Machine Learning Research*, 21(1):5709–5738, 2020.

Pascal Massart. *Concentration Inequalities and Model Selection*, volume 6. Springer, 2007.

Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society Series B*, 79(4): 1119–1141, 2017.

Daniel A McFarland. Student resistance: How the formal and informal organization of classrooms facilitate everyday forms of student defiance. *American Journal of Sociology*, 107(3):612–678, 2001.

Mark Newman. *Networks*. Oxford university press, 2018.

Oscar Hernan Madrid Padilla, James Sharpnack, and James G Scott. The dfs fused lasso: Linear-time denoising over general graphs. *The Journal of Machine Learning Research*, 18(1):6410–6445, 2017.

Debdeep Pati, Anirban Bhattacharya, and Yun Yang. On statistical optimality of variational Bayes. In *International Conference on Artificial Intelligence and Statistics*, pages 1579–1588. PMLR, 2018.

Judea Pearl. *Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach*. Cognitive Systems Laboratory, School of Engineering and Applied Science, 1982.

Marianna Pensky. Dynamic network models and graphon estimation. *The Annals of Statistics*, 47(4):2378–2403, 2019.

Nicholas G Polson and James G Scott. On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012.

Purnamrita Sarkar and Andrew W Moore. Dynamic social network analysis using latent space models. *Acm Sigkdd Explorations Newsletter*, 7(2):31–40, 2005.

Daniel K Sewell and Yuguo Chen. Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110(512):1646–1657, 2015.

Daniel K Sewell and Yuguo Chen. Latent space approaches to community detection in dynamic networks. *Bayesian Analysis*, 12(2):351–377, 2017.

Qi-Man Shao. A note on small ball probability of a Gaussian process with stationary increments. *Journal of Theoretical Probability*, 6(3):595–602, 1993.

Tom AB Snijders. Statistical models for social networks. *Annual Review of Sociology*, 37: 131–153, 2011.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

Aad W Van der Vaart and J Harry Van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

Vincent Vu and Jing Lei. Minimax rates of estimation for sparse PCA in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 1278–1286. PMLR, 2012.

Martin J Wainwright and Michael Irwin Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc, 2008.

Stephen Walker and Nils Lid Hjort. On Bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):811–821, 2001.

Bo Wang and DM Titterington. Lack of consistency of mean field and variational Bayes approximations for state space models. *Neural Processing Letters*, 20(3):151–170, 2004.

Yixin Wang and David M Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

Yair Weiss and Judea Pearl. Belief propagation: technical perspective. *Communications of the ACM*, 53(10):94–94, 2010.

Eric P Xing, Wenjie Fu, and Le Song. A state-space mixed membership blockmodel for dynamic network tomography. *Annals of Applied Statistics*, 4(2):535–566, 2010.

Kevin S Xu and Alfred O Hero. Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):552–562, 2014.

Tianbao Yang, Yun Chi, Shenghuo Zhu, Yihong Gong, and Rong Jin. Detecting communities and their evolutions in dynamic social networks—a Bayesian approach. *Machine Learning*, 82(2):157–189, 2011.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905, 2020.

Anderson Y Zhang and Harrison H Zhou. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016.

Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180–2207, 2020.

Jingnan Zhang, Xin He, and Junhui Wang. Directed community detection with network embedding. *Journal of the American Statistical Association*, 117(540):1809–1819, 2022a.

Xuefei Zhang, Songkai Xue, and Ji Zhu. A flexible latent space model for multilayer networks. In *International Conference on Machine Learning*, pages 11288–11297. PMLR, 2020.

Xuefei Zhang, Gongjun Xu, and Ji Zhu. Joint latent space models for network data with high-dimensional node variables. *Biometrika*, 109(3):707–720, 2022b.

Peng Zhao, Anirban Bhattacharya, Debdeep Pati, and Bani K Mallick. Factorized fusion shrinkage for dynamic relational data. *arXiv preprint arXiv:2210.00091*, 2022.