High-Dimensional Differential Parameter Inference in Exponential Family using Time Score Matching

Daniel J. Williams[*,1] **Leyang Wang**[*,1] **Qizhen Ying**[1]
Song Liu[1] **Mladen Kolar**[2]
[1]School of Mathematics, University of Bristol [2]USC Marshall School of Business

Abstract

This paper addresses differential inference in time-varying parametric probabilistic models, like graphical models with changing structures. Instead of estimating a high-dimensional model at each time point and estimating changes later, we directly learn the differential parameter, i.e., the time derivative of the parameter. The main idea is treating the time score function of an exponential family model as a linear model of the differential parameter for direct estimation. We use time score matching to estimate parameter derivatives. We prove the consistency of a regularized score matching objective and demonstrate the finite-sample normality of a debiased estimator in high-dimensional settings. Our methodology effectively infers differential structures in high-dimensional graphical models, verified on simulated and real-world datasets. The code reproducing our experiments can be found at: `https://github.com/Leyangw/tsm`.

1 Introduction

In non-stationary environments, the data-generating process varies over time due to factors like news, geopolitical events, and economic reports (Lu et al., 2019; Quiñonero-Candela et al., 2022). Understanding these changes is crucial for many applications.

When probabilistic model parameters change over time, learning their time-derivative, or differential parameter, can be beneficial. Although time-varying models are well-studied (Kolar and Xing, 2011, 2012; Gibberd and Nelson, 2017; Yang and Peng, 2020), few focus on differential parameters. Learning these is advantageous: they reveal the underlying dynamics of systems, such as the SIR model (Tang et al., 2020). Differential parameters can also be more interpretable; stationary parameters become zero after differentiation, making the estimation target *sparse*. This helps in handling high-dimensional overparametrized models (Hastie et al., 2015; Wainwright, 2019) and inferring with fewer samples.

Differential model estimation has been considered in a "discrete setting" (Zhao et al., 2014; Liu et al., 2017; Zhao et al., 2019; Kim et al., 2021), where the focus is on identifying changes in the model between two discrete time points. However, in this paper, we learn the differential parameter in a continuous setting, where the parameter is a continuous function of the time.

We propose an efficient estimator for the differential parameter in high-dimensional probabilistic models. Our method estimates the differential parameter directly, without needing to estimate the parameters themselves. By addressing an ℓ_1-regularized objective, our estimator achieves consistency in high-dimensional contexts, with a convergence rate dependent only on the dimensionality of the time-varying parameters. The debiased estimator shows asymptotic normality, making it suitable for parameter inference, and it efficiently estimates complex models without evaluating the normalizing term. We validate our theorems through synthetic experiments and demonstrate superior performance compared to a recent time-varying parameter estimation method (Yang and Peng, 2020). This is the first work to tackle the differential parameter estimation in exponential family in a continuous setting.

2 Background

Let $\boldsymbol{x} \in \mathbb{R}^d$ be a random sample generated from a distribution whose density function is $q(\boldsymbol{x})$. We also

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

define a parametric model $p(\boldsymbol{x};\boldsymbol{\theta})$

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\bar{p}(\boldsymbol{x};\boldsymbol{\theta})}{z(\boldsymbol{\theta})}, \;\; z(\boldsymbol{\theta}) = \int \bar{p}(\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x}, \qquad (1)$$

where $\bar{p}(\boldsymbol{x};\boldsymbol{\theta})$ is known analytically and $z(\boldsymbol{\theta})$ is the normalizing constant which may be hard to compute or approximate.

Before introducing the proposed approach, we first review two important notions, high dimensional probabilistic graphical model and score matching.

2.1 Estimating Probabilistic Model Parameters in High-dimensional Settings

Suppose we are interested in using samples drawn from q to learn an over-parametrized model $p(\boldsymbol{x};\boldsymbol{\theta})$ with more parameters than necessary to describe the data. The "true model" generating data might be simpler with few parameters. Sparse probabilistic model estimation aims to find a *sparse parameterization* of the high-dimensional model that best fits the data. Previous studies have used sparsity-inducing regularization (e.g., ℓ_1 penalty) along with the likelihood function on the dataset $\mathcal{D}$ (Tibshirani, 1996; Hastie et al., 2015; Yuan and Lin, 2007; Drton and Maathuis, 2017). Consider a pairwise graphical model

$$\bar{p}(\boldsymbol{x};\boldsymbol{\Theta}) := \exp\left(\sum_{i\leq j} \Theta_{i,j} f(x_i, x_j)\right), i, j \in [d].$$

A large d results in $\boldsymbol{\Theta}$ being high-dimensional, leading to $p(\boldsymbol{x};\boldsymbol{\Theta})$ being over-parameterized. Assuming the true model $q(\boldsymbol{x}) = p(\boldsymbol{x};\boldsymbol{\Theta}^*)$ is sparse with parameter $\boldsymbol{\Theta}^*$, the regularized likelihood estimator can effectively estimate the graphical model's parameter.

2.2 Score Matching (Hyvärinen, 2005)

Score matching involves the score functions of the data which are independent of the normalising constant $z(\boldsymbol{\theta})$. The classical score matching objective involves taking the expected squared distance between target score $\nabla_{\boldsymbol{x}} \log q(\boldsymbol{x})$ and model score $\nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$, given by

$$\mathcal{L}_{\mathrm{SM}}(\boldsymbol{\theta}) = \mathbb{E}_q \left[\|\nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x})\|^2\right]. \quad (2)$$

Note that the Equation (2) is not tractable due to the unknown $\nabla_{\boldsymbol{x}} \log q(\boldsymbol{x})$. Hyvärinen (2005) showed that integration by parts enables a tractable version of Equation (2) to eliminate any unknown quantities.

Score matching has been extended for numerous settings, which include *generalised* score matching (Hyvärinen, 2007; Yu et al., 2019), applications to high-dimensional graphical models Lin et al. (2016a); Yu et al. (2016) and truncated density estimation (Liu et al., 2022; Williams and Liu, 2022).

3 Formulation and Motivation

In this section, we formally introduce the problem of estimating differential parameters in exponential family distributions. Let's assume that the true time-varying data generating distribution has a density function $q_t(\boldsymbol{x})$. $q_t(\boldsymbol{x})$ is in the exponential family and is parameterized as:

$$q_t(\boldsymbol{x}) = q(\boldsymbol{x};\boldsymbol{\theta}^*(t)) = \frac{\exp\left(\langle \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{x})\rangle\right)}{z(\boldsymbol{\theta}^*(t))}, \qquad (3)$$

where the normalising term $z(\boldsymbol{\theta}^*(t))$ is defined as

$$z(\boldsymbol{\theta}^*(t)) := \int \exp\left(\langle \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{x})\rangle\right) \mathrm{d}\boldsymbol{x},$$

which is generally computationally intractable. The density function has a natural parameter $\boldsymbol{\theta}^*(t)$ that changes with t. We assume that $\partial_t \boldsymbol{\theta}^*(t)$ is a sparse vector, i.e., only a few elements in $\boldsymbol{\theta}^*$ depend on t. We denote $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^k$ as the *feature function*, which does not change over time.

3.1 Learning Differential Parameters, not Parameters

In this paper, we want to know how q_t changes over time. Specifically, we want to learn the time derivative of the parameter, i.e., $\partial_t \boldsymbol{\theta}^*(t)$.

We could adopt the approach that approximates the density function $q(\boldsymbol{x};\boldsymbol{\theta}^*(t))$ with a parametric model $p(\boldsymbol{x};\hat{\boldsymbol{\theta}}(t))$ and subsequently, differentiates the estimated parameter $\hat{\boldsymbol{\theta}}(t)$ with respect to t. However, if only a few elements in $\boldsymbol{\theta}^*(t)$ change with t, then modelling and learning the full density, including all stationary parameters that do not change with t, is inefficient, especially when $\boldsymbol{\theta}^*(t)$ is high dimensional.

3.2 Modeling Time Score via Differential Parameters

In the remainder of this paper, we refer to $\partial_t \log q_t(\boldsymbol{x})$, the time derivative of the log density, as the "time score function".

A key observation that motivated this work is that the time score function can be expressed as a function of the differential parameter $\partial_t \boldsymbol{\theta}(t)$. It follows from Equation (3) that $\partial_t \log q_t(\boldsymbol{x})$ has a simple closed-form expression given in the following proposition.

Proposition 3.1. *The time score function for q_t, which is defined in Equation* (3)*, can be written as*

$$\partial_t \log q_t(\boldsymbol{x}) = \langle \partial_t \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{x})\rangle - \mathbb{E}_{\boldsymbol{y}\sim q_t}\left[\langle \partial_t \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{y})\rangle\right].$$

A proof is given in Appendix B. Proposition 3.1 shows that $\partial_t \log q_t(\boldsymbol{x})$ does not involve the normalization constant explictly and depends solely on the differential

parameter $\partial_t \boldsymbol{\theta}^*(t)$. This is an analogue to how the density ratio function only depends on the difference between two parameters $\boldsymbol{\theta}_0 - \boldsymbol{\theta}_1$ as we explain in Section 7.2. Inspired by Proposition 3.1, we directly model the time score function $\partial_t \log q_t(\boldsymbol{x})$ as

$$s(\boldsymbol{x}, t) \coloneqq \langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x}) \rangle - \mathbb{E}_{\boldsymbol{y} \sim q_t} \left[\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y}) \rangle \right], \quad (4)$$

where $\partial_t \boldsymbol{\theta}(t)$ is a parameterization of the differential parameter that we will detail later. If we can estimate the model in Equation (4), we obtain the differential parameter $\partial_t \boldsymbol{\theta}(t)$.

4 Estimator of Differential Parameters

4.1 Time Score Matching

Our method seeks to approximate the time score function $\partial_t \log q_t(\boldsymbol{x})$ by $s(\boldsymbol{x}, t)$. We adopt a learning criterion called time score matching (Choi et al., 2022). Specifically, we look for $\partial_t \boldsymbol{\theta}(t)$ that minimizes the weighted Fisher-Hyvärinen divergence (Lyu, 2009)

$$\mathcal{L}(\partial_t \boldsymbol{\theta}(t)) \coloneqq \int_{t=0}^{t=1} \mathbb{E}_{q_t} \left[g(t) \| \partial_t \log q_t(\boldsymbol{x}) - s(\boldsymbol{x}, t) \|^2 \right] dt. \quad (5)$$

Choi et al. (2022) uses a neural network to model the time score, which does not allow the direct inference of $\partial_t \boldsymbol{\theta}(t)$. However, in our paper, Proposition 3.1 relates $\partial_t \boldsymbol{\theta}(t)$ to the time score function, enabling the direct estimation of $\partial_t \boldsymbol{\theta}(t)$ using time score matching.

Equation (5) is a truncated score matching problem (Liu et al., 2022; Yu et al., 2022): The variable being differentiated in the log-density function, t is in a bounded domain $[0, 1]$. The following theorem provides a tractable learning objective without intractable terms such as $\partial_t \log q_t(\boldsymbol{x})$.

Theorem 4.1. *Suppose that $g(t) = 0$ for $t = 0$ or 1. The objective in (5) can be written as*

$$\mathcal{L}(\partial_t \boldsymbol{\theta}(t)) = C + \int_0^1 \mathbb{E}_{q_t} \left[g(t) s(\boldsymbol{x}, t)^2 \right] \mathrm{d}t + \int_0^1 2\mathbb{E}_{q_t} \left[\partial_t g(t) \langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x}) \rangle + g(t) \left\langle \partial_t^2 \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x}) \right\rangle \right] \mathrm{d}t \quad (6)$$

where C is a constant independent of $s(\boldsymbol{x}; t)$.

A proof is given in Appendix C. Theorem 4.1 is *not* a starightforward application of Theorem 1 in (Choi et al., 2022). The $\mathbb{E}_{\boldsymbol{y} \sim q_t} \left[\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y}) \rangle \right]$ term of our score model requires additional steps after the initial integration by parts and the resulting objective is different from Equation (8) in (Choi et al., 2022). Noticing that the weighting technique introduced in (Liu et al., 2022; Yu et al., 2022), we let g be zero at the boundary $t = 0$ and 1 to ensure the boundary condition used in integration by parts is satisfied. Moreover, in (Choi et al., 2022), authors convert Equation (5) into a different form, which requires multiple samples of $\boldsymbol{x}$ at $t = 0$ and $t = 1$ for Monte Carlo approximation. However, this is not a requirement for our method. The exact form of g can be found in the Appendix G.3.

Note that this objective function only depends on $\partial_t \boldsymbol{\theta}(t)$, so no nuisance parameters are required when optimizing the above objective. To finally transform Equation (6) into a tractable optimization problem, we need to approximate it using samples and design a parametric model for $\partial_t \boldsymbol{\theta}(t)$.

4.2 Sample Approximation of Objective Function

In this section, we consider sample approximations of Equation (6) under two different settings. In the first scenario, we assume that we have paired samples from the joint distribution of both $\boldsymbol{x}$ and t, i.e.

$$\mathcal{D}_1 \coloneqq \{(t_i, \boldsymbol{x}_i)\}_{i=1}^n \sim q(\boldsymbol{x}, t),$$

where $q(\boldsymbol{x}, t) = q_t(\boldsymbol{x}) \times \mathrm{Uniform}(t; 0, 1)$, where we shortened $q(\boldsymbol{x}|t)$ as $q_t(\boldsymbol{x})$ and the same below. Then the sample objective of Equation (6) (omitting the constant) can be written as

$$\frac{1}{n} \sum_{i=1}^n g(t_i) s(\boldsymbol{x}_i, t_i)^2 + 2\partial_t g(t_i) \langle \partial_t \boldsymbol{\theta}(t_i), \boldsymbol{f}(\boldsymbol{x}_i) \rangle + 2g(t_i) \left\langle \partial_t^2 \boldsymbol{\theta}(t_i), \boldsymbol{f}(\boldsymbol{x}_i) \right\rangle. \quad (7)$$

In the second scenario, we assume that we first draw samples $\{t_j\}_{j=1}^m \sim \mathrm{Uniform}(t; 0, 1)$, then draw samples $\{\boldsymbol{x}_{ij}\}_{i=1}^n \sim q_{t_j}(\boldsymbol{x})$ for each $j \in [m]$. Given the dataset

$$\mathcal{D}_2 \coloneqq \{(t_j, \{\boldsymbol{x}_{ij}\}_{i=1}^n)\}_{j=1}^m,$$

the sample objective of Equation (6) (omitting the constant) is written as

$$\frac{1}{m} \sum_{j=1}^m \frac{1}{n} \sum_{i=1}^n g(t_j) s(\boldsymbol{x}_{ij}, t_j)^2 + 2\partial_t g(t_j) \langle \partial_t \boldsymbol{\theta}(t_j), \boldsymbol{f}(\boldsymbol{x}_{ij}) \rangle + 2g(t_j) \left\langle \partial_t^2 \boldsymbol{\theta}(t_j), \boldsymbol{f}(\boldsymbol{x}_{ij}) \right\rangle. \quad (8)$$

Both scenarios naturally arise in machine learning tasks. The first scenario resembles a time series setting, where we have a single sample for each time point. In contrast, the second scenario aligns more with a continuous dataset drift setting (Quiñonero-Candela et al., 2022), where we receive a full set of samples for each time point.

4.3 Sample Approximation of Score Model

Now we look at the score model used in above objectives

$$s(\boldsymbol{x}, t) := \langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x}) \rangle - \mathbb{E}_{\boldsymbol{y} \sim q_t} \left[\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y}) \rangle \right].$$

From the definition, we can see that $s(\boldsymbol{x}, t)$ contains term $\mathbb{E}_{q_t} \left[\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x}) \rangle \right]$, which is an *expectation conditioned on* t. In general, this time-conditional expectation does not have a closed-form expression. Thus, we also have to approximate it using samples.

In the first scenario, where we only have access to a paired dataset $\mathcal{D}_1$, we consider using **Nadarya-Watson (NW) estimator** (Nadaraya, 1964; Watson, 1964) to estimate the conditional expectation. NW estimator is a weighted sample average. In our context, we can write the NW estimator as

$$\hat{\mathbb{E}}_{q_{t_i}} \left[\langle \partial_t \boldsymbol{\theta}(t_i), \boldsymbol{f}(\boldsymbol{y}) \rangle \right] := \frac{\sum_{j=1}^m K(t_j, t_i) \langle \partial_t \boldsymbol{\theta}(t_j), \boldsymbol{f}(\boldsymbol{x}_j) \rangle}{\sum_{j=1}^m K(t_j, t_i)}, \tag{9}$$

where K is a Gaussian kernel function.

In the second scenario, given the dataset $\mathcal{D}_2$, we can simply approximate the expectation as the average of samples obtained at time point t_j:

$$\hat{\mathbb{E}}_{q_{t_j}} \left[\langle \partial_t \boldsymbol{\theta}(t_j), \boldsymbol{f}(\boldsymbol{y}) \rangle \right] := \frac{1}{n} \sum_{i=1}^n \langle \partial_t \boldsymbol{\theta}(t_j), \boldsymbol{f}(\boldsymbol{x}_{ij}) \rangle, \forall j \in [m].$$

Now we denote the approximated score model as

$$\hat{s}(\boldsymbol{x}, t) := \langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x}) \rangle - \hat{\mathbb{E}}_{q_t} \left[\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y}) \rangle \right],$$

where $\hat{\mathbb{E}}_{q_t} \left[\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y}) \rangle \right]$ is the approximated conditional expectation in the above two settings.

4.4 Parameterization of $\partial_t \boldsymbol{\theta}(t)$

So far, the objective function in Equation (6) works for any generic parameterisation of $\partial_t \boldsymbol{\theta}(t)$. Without the loss of generality, we propose a parametric model

$$\partial_t \boldsymbol{\theta}(t) := [\partial_t \theta_1(t), \ldots, \partial_t \theta_k(t)]^\top, \partial_t \theta_j(t) := \langle \boldsymbol{\alpha}_j, \partial_t \boldsymbol{\phi}(t) \rangle,$$

where $\boldsymbol{\phi} : \mathbb{R} \to \mathbb{R}^b$ is a differentiable basis function and $\boldsymbol{\alpha}_j \in \mathbb{R}^b$ is a parameter vector to be estimated.

In the simplest case, $\phi(t) := t$, so that $\partial_t \theta_j(t) := \alpha_j$. This model implies that $\boldsymbol{\theta}(t)$ changes with t *linearly*. One can consider other basis functions, for example, the Fourier basis, which are often used to model time-dependent functions,

$$\boldsymbol{\phi}(t) = [\sin(t), \cos(t), \cdots, \sin(bt/2), \cos(bt/2)].$$

Suppose we adopt a linear model, i.e., $\phi(t) := t$. We can rewrite the time score model as

$$\hat{s}_{\boldsymbol{\alpha}}(\boldsymbol{x}, t) := \langle \boldsymbol{\alpha}, \boldsymbol{f}(\boldsymbol{x}) \rangle - \hat{\mathbb{E}}_{\boldsymbol{y} \sim q_t} \left[\langle \boldsymbol{\alpha}, \boldsymbol{f}(\boldsymbol{y}) \rangle \right].$$

Thus, given the dataset $\mathcal{D}_1$, we have the following tractable objective:

$$\hat{\mathcal{L}}(\boldsymbol{\alpha}) := \frac{1}{n} \sum_{i=1}^n g(t_i) \hat{s}_{\boldsymbol{\alpha}}(\boldsymbol{x}_i; t_i)^2 + 2 \partial_t g(t_i) \langle \boldsymbol{\alpha}, \boldsymbol{f}(\boldsymbol{x}_i) \rangle, \tag{10}$$

where we have replaced s in Equation (7) with $\hat{s}_{\boldsymbol{\alpha}}$ and $\partial_t \boldsymbol{\theta}(t)$ with the parameter $\boldsymbol{\alpha}$. Note that the third term in (7) vanishes since $\partial_t^2 \boldsymbol{\theta}(t) = 0$ given the choice $\phi(t) = t$.

Define $\boldsymbol{F} \in \mathbb{R}^{n \times k}$ as the feature matrix whose i-th row is $\boldsymbol{f}(\boldsymbol{x}_i)$, centered sufficient statistic $\tilde{\boldsymbol{F}} := \boldsymbol{F} - \hat{\mathbb{E}}_{q_t}[\boldsymbol{F}]$, and a diagonal matrix $\boldsymbol{G} \in \mathbb{R}^{n \times n}$ whose i-th diagonal entry is $g(t_i)$. We can rewrite Equation (10) using the following equivalent quadratic form

$$\hat{\mathcal{L}}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \tilde{\boldsymbol{F}}^\top \boldsymbol{G} \tilde{\boldsymbol{F}} \boldsymbol{\alpha} / n + 2 \mathbf{1}_n^\top \partial_t \boldsymbol{G} \boldsymbol{F}^\top \boldsymbol{\alpha} / n, \tag{11}$$

and the minimizer of the above objective has a closed-form expression $\left[\tilde{\boldsymbol{F}}^\top \boldsymbol{G} \tilde{\boldsymbol{F}} \right]^{-1} \boldsymbol{F}^\top \partial_t \boldsymbol{G} \mathbf{1}_n$. However, in the high dimensional setting, where the number of samples n is potentially smaller than the dimension of both $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}(t)$, $\tilde{\boldsymbol{F}}^\top \boldsymbol{G} \tilde{\boldsymbol{F}}$ is non-invertible. We introduce a high-dimensional estimator of $\boldsymbol{\alpha}$.

5 High-dimensional Differential Parameter Estimation and Debiasing

To simplify our discussion, from now on, we suppose that we have access to dataset $\mathcal{D}_1$ and $\phi(t) = t$.

In high-dimensional settings, we assume only a few elements in $\boldsymbol{\theta}^*(t)$ change over time t, making $\partial_t \boldsymbol{\theta}^*(t)$ (and $\boldsymbol{\alpha}$) a sparse vector. Hence, we use a lasso regularizer to identify the non-zeros in $\partial_t \boldsymbol{\theta}(t)$, where we refer to this lasso estimator of $\hat{\boldsymbol{\alpha}}$ as "SparTSM". We propose minimizing $\hat{\mathcal{L}}(\boldsymbol{\alpha})$ with a sparsity-inducing ℓ_1 norm:

$$\hat{\boldsymbol{\alpha}} := \operatorname*{argmin}_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}) + \lambda_{\text{lasso}} \|\boldsymbol{\alpha}\|_1. \tag{12}$$

In Section 6, we prove the consistency of $\hat{\boldsymbol{\alpha}}$ in a high dimensional setting. Using a lasso estimator in Equation (12) can introduce biases, making the asymptotic distribution of $\hat{\boldsymbol{\alpha}}$ intractable. This is not ideal if we are interested in parameter inference, such as hypothesis tests and establishing confidence intervals. We apply the debiasing technique (Zhang and Zhang, 2014; van de Geer et al., 2014) to the lasso estimate of each component, which will allow us to track the asymptotic distribution. Let $\boldsymbol{\omega}_j^*$ denote the j-th column of the inverse Hessian $\left[\nabla^2_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^*) \right]^{-1}$ and $\tilde{\boldsymbol{\omega}}_j$ be a consistent estimator of $\boldsymbol{\omega}_j^*$. We debias the j-th element of the lasso estimate using a single-step Newton update:

$$\tilde{\boldsymbol{\alpha}}_j = \hat{\boldsymbol{\alpha}}_j - \tilde{\boldsymbol{\omega}}_j^\top \nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\hat{\boldsymbol{\alpha}}). \tag{13}$$

Algorithm 1 Inference Pipeline

Require: Dataset: $\{(t_i, \boldsymbol{x}_i)\}_{i=1}^n$,
Regularization parameters $\lambda_{\text{lasso}}, \lambda_1, ..., \lambda_k > 0$.
Find lasso estimator $\hat{\boldsymbol{\alpha}}$ by solving (12)
for $j \in [k]$ **do**
  Find $\tilde{\boldsymbol{\omega}}_j$ by solving (14)
  Obtain debiased lasso by (13).
end for
return asymptotically unbiased estimator $\tilde{\boldsymbol{\alpha}}$

Estimating the inverse Hessian $\left[\nabla^2_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^*)\right]^{-1}$ in high-dimensional space is challenging, as the empirical Hessian $\nabla^2_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\hat{\boldsymbol{\alpha}})$ is ill-conditioned and often non-invertible. Fortunately, $\boldsymbol{\omega}_j^*$ satisfies the equality $\left[\nabla^2_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^*)\right] \boldsymbol{\omega}_j^* = \boldsymbol{e}_j$, where $\boldsymbol{e}_j$ is a vector with the j-th element equal to one and zeros elsewhere. We estimate $\tilde{\boldsymbol{\omega}}_j$ using the ℓ_1-norm regularized objective:

$$\tilde{\boldsymbol{\omega}}_j = \arg\min_{\boldsymbol{\omega}} \frac{1}{2} \boldsymbol{\omega}^\top \nabla^2_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\hat{\boldsymbol{\alpha}}) \boldsymbol{\omega} - \boldsymbol{\omega}^\top \boldsymbol{e}_j + \lambda_j ||\boldsymbol{\omega}||_1 \quad (14)$$

The consistency of this estimator has been proved in Appendix E.2. We show in Section 6 that the debiased estimator in Equation (13) is asymptotically unbiased and normally distributed under further conditions. We summarize the high-dimensional differential parameter inference pipeline in Algorithm 1. We refer to the estimator of $\tilde{\boldsymbol{\alpha}}_j$ in Equation (13) as "SparTSM+"

6 Theoretical Analysis

We show that both SparTSM and SparTSM+ work effectively in a high-dimensional regime when $\partial_t \boldsymbol{\theta}(t)$ is sparse. A full list of notations is provided in Appendix A. For our theoretical results, we assume the following:

Assumption 6.1. There exists a unique parameter $\boldsymbol{\alpha}^*$ supported on $S \subseteq \{1, \ldots, k\}$ such that $\partial_t \boldsymbol{\theta}_t^* = \boldsymbol{\alpha}^*$, and the population objective $\mathcal{L}(\boldsymbol{\alpha})$ is minimised at $\boldsymbol{\alpha}^*$.

6.1 Finite-sample Estimation Error of SparTSM

We assume bounded sufficient statistics in the exponential family model (Kim et al., 2021; Xia et al., 2023).

Assumption 6.2. There exists some constant $0 < C_f < \infty$ such that $\|\boldsymbol{f}(\boldsymbol{x})\|_\infty \leq C_f$ almost surely.

Defining $\boldsymbol{G}^{1/2} \in \mathbb{R}^{n \times n}$ as a diagonal matrix where $G^{1/2}_{i,i} = \sqrt{g(t_i)}$, we assume the following:

Assumption 6.3. The matrix $\boldsymbol{G}^{1/2} \tilde{\boldsymbol{F}}$ satisfies the restricted eigenvalue (RE) condition over the support set S with parameters $(\kappa, 3)$, that is,

$$\frac{1}{n} \|\boldsymbol{G}^{1/2} \tilde{\boldsymbol{F}} \Delta\|_2^2 \geq \kappa \|\Delta\|_2^2 \text{ for all } \Delta \in \mathbb{C}_3(S), \quad (15)$$

where $\mathbb{C}_3(S) = \{\Delta \in \mathbb{R}^k : \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\}$.

Building on the aforementioned assumption, we can establish a probabilistic upper bound for the ℓ_2 norm of the error vector $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*$.

Define a random variable

$$m_{\boldsymbol{\alpha}}(\boldsymbol{x}, t) := g(t) \hat{s}_{\boldsymbol{\alpha}}(\boldsymbol{x}; t)^2 + 2 \partial_t g(t) \langle \boldsymbol{\alpha}, \boldsymbol{f}(\boldsymbol{x}) \rangle$$

where $(\boldsymbol{x}, t) \sim q_t(\boldsymbol{x}) \times \text{Uniform}(t; 0, 1)$.

Theorem 6.4. *Suppose Assumption 6.1,6.2 and 6.3 hold. Let*

$$\sigma^2 = \max_{1 \leq i \leq k} \boldsymbol{\Sigma}_{ii}(\boldsymbol{\alpha}^*), \quad \boldsymbol{\Sigma}(\boldsymbol{\alpha}) := \text{Cov}[\nabla_{\boldsymbol{\alpha}} m_{\boldsymbol{\alpha}}(\boldsymbol{x}, t)],$$

and we set $\lambda_{\text{lasso}} = 2\sigma \left(\sqrt{\frac{2 \log k}{n}} + \delta \right)$. *It holds that*

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2 \leq \frac{6}{\kappa} \|\boldsymbol{\alpha}^*\|_0^{1/2} \sigma \left(\sqrt{\frac{2 \log k}{n}} + \delta \right) \text{ for all } \delta > 0. \quad (16)$$

with probability at least $1 - 2\exp\{-\frac{n\delta^2}{\sigma}\}$.

See Appendix D.2 for the proof.

Remarks Crucially, the error bound only depends on the dimension of the feature function k logarithmically, indicating the estimator indeed scales to high dimensional settings. Moreover, the error bound depends on the sparsity of $\boldsymbol{\alpha}^*$ (i.e., the sparsity of $\partial_t \boldsymbol{\theta}^*(t)$) and does *not* depend on the sparsity of $\boldsymbol{\theta}(t)$. It means that our method works with a dense parameter vector $\boldsymbol{\theta}(t)$ as long as $\partial_t \boldsymbol{\theta}(t)$ is sparse. This opens the door for applications where time-varying probabilistic models that are complex and cannot be described by sparse models. We showed an example in Section 8.

6.2 Finite-sample Gaussian Approximation Bound of SparTSM+

In this section, we prove that $\sqrt{n}\left(\tilde{\alpha}_j - \alpha_j^*\right)/\hat{\sigma}_j$ converges to a standard normal random variable. where $\hat{\sigma}_j$ is defined as

$$\hat{\sigma}_j^2 = \tilde{\boldsymbol{\omega}}_j^\top \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}) \tilde{\boldsymbol{\omega}}_j, \quad \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}) = \text{Cov}_n[\nabla_{\boldsymbol{\alpha}} m_{\hat{\boldsymbol{\alpha}}}(\boldsymbol{x}, t)], \quad (17)$$

and Cov_n is the empirical covariance estimator. Now we justify the asymptotic normality using the following Gaussian Approximation Bound (GAB), showing that as the number of samples increases, the distribution of $\tilde{\alpha}_j$ is closing to a Gaussian distribution.

Theorem 6.5 (GAB). *Suppose Assumption 6.1 and 6.2 hold. For $\delta_{\boldsymbol{\alpha}}, \delta_w, \lambda_{\text{lasso}}, \lambda_j, \delta_\sigma \in [0,1)$, define the set of events $\boldsymbol{\mathcal{E}}$,*

$$\boldsymbol{\mathcal{E}} = \begin{cases} \|\nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\|_\infty \le \lambda_{\text{lasso}}/2; \\ \|\nabla^2_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\boldsymbol{\omega}_j^* - \boldsymbol{e}_j\|_\infty \le \lambda_j/2; \\ \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 \le \delta_{\boldsymbol{\alpha}}; \\ \|\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\|_1 \le \delta_w; \\ ||\hat{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}^*) - \boldsymbol{\Sigma}||_\infty \le \delta_\sigma/2. \end{cases}$$

Suppose $\mathbb{P}(\boldsymbol{\mathcal{E}}) \ge 1-\epsilon$, denoting $\Phi(\cdot)$ as the cumulative distribution function of the standard normal distribution, we have

$$\sup_{z\in\mathbb{R}} \left|\mathbb{P}\left\{\sqrt{n}(\tilde{\alpha}_j - \alpha_j^*)/\hat{\sigma}_j \le z\right\} - \Phi(z)\right| \le \Delta_1 + \Delta_2 + \frac{\delta_C}{1-\delta_C} + \epsilon \quad (18)$$

where

$$\Delta_1 = 2CM||\boldsymbol{\omega}_j^*||_1/\sqrt{n}\sigma_j,$$
$$\Delta_2 = \sqrt{n}(\lambda_j\delta_{\boldsymbol{\alpha}} + \delta_w\lambda_{\text{lasso}} + K\delta_{\boldsymbol{\alpha}}\delta_w)/\boldsymbol{\omega}_j^*\boldsymbol{\Sigma}(\boldsymbol{\alpha}^*)\boldsymbol{\omega}_j^*,$$
$$\delta_C = ((2L\delta_\alpha + \delta_\sigma)(||\boldsymbol{\omega}_j^*||_1^2 + \delta_w^2) + |||\boldsymbol{\Sigma}|||_\infty\delta_w^2)/\sigma_j^2$$

here, C, M, K, L are fixed constants.

See Appendix E.3 for proof. The proof involves decomposing the standardized debiased estimator into three parts. One part introduces the desired asymptotic normality due to the Berry-Esseen inequality (Chen et al., 2010), while the other two converge to zero conditioned on $\boldsymbol{\mathcal{E}}$.

Assumption 6.6. The matrix $\boldsymbol{G}^{1/2}\tilde{\boldsymbol{F}}$ satisfies the RE condition over the support set $S_{\boldsymbol{\omega},j}$ with parameters $(\kappa_j, 3)$ where $S_{\boldsymbol{\omega},j}$ is the support set of $\boldsymbol{\omega}_j^*$.

We can further specify the rate of the approximation under appropriate settings of λ_{lasso} and λ_j.

Corollary 6.7. *Assume 6.1,6.2, 6.3 and 6.6 hold. Let $\tilde{\alpha}_j$ be the debiased lasso estimator derived by Equation* (13) *with the regularization parameters set as*

$$\lambda_{\text{lasso}} \in \mathcal{O}\left(\sqrt{\frac{\log k}{n}}\right),\ \lambda_j \in \mathcal{O}\left(\sqrt{\|\boldsymbol{\omega}_j^*\|_0 \frac{\log k}{n}}\right) \quad (19)$$

then there exist positive constants c, c' such that

$$\sup_{z\in\mathbb{R}} \left|\mathbb{P}\{\sqrt{n}(\tilde{\alpha}_j - \alpha_j^*)/\hat{\sigma}_j \le z\} - \Phi(z)\right| \le \mathcal{O}\left(\|\boldsymbol{\omega}_j^*\|_0^{3/2}\|\boldsymbol{\alpha}^*\|_0 \frac{\log k}{\sqrt{n}}\right) + c\exp\{-c'\log k\} \quad (20)$$

See Appendix E.4 for the proof.

Remarks This result shows that, as sample size increases, the standardized α_j indeed converges to a standard normal variable, allowing us to specify the confidence interval and perform hypothesis tests. Similar to the error bound stated in Theorem 6.4, this approximation error bound also only depends on the sparsity of $\boldsymbol{\alpha}^*$ (i.e., the sparsity of $\partial_t\boldsymbol{\theta}^*(t)$), rather than the sparsity of $\boldsymbol{\theta}^*(t)$, indicating it is applicable to complex probabilistic models with only sparse changes. We also notice that the result depends on the sparsity of the inverse Hessian column vector, $\boldsymbol{\omega}_j^*$, which is similar to the previous debiased lasso results (Kim et al., 2021).

7 Related Works

We now introduce two additional methods for learning the differential parameters in graphical models, which are later compared in the simulation studies.

7.1 Estimating Parameter Changes in Probabilistic Models

Given two data-generating distributions, $q_0(\boldsymbol{x}) = p(\boldsymbol{x};\boldsymbol{\theta}_0^*)$ and $q_1(\boldsymbol{x}) = p(\boldsymbol{x};\boldsymbol{\theta}_1^*)$, we are interested in learning changes in the underlying data-generating distributions, given random samples $\boldsymbol{x}_0 \sim q_0$ and $\boldsymbol{x}_1 \sim q_1$.

One naive way of estimating the parameter change is fitting two probabilistic models $p(\boldsymbol{x};\boldsymbol{\theta}_0)$ and $p(\boldsymbol{x};\boldsymbol{\theta}_1)$ from $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ using lasso estimators, then take the difference of estimated parameters $\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0$. This approach is sub-optimal, as sparse estimates of $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_0$ do not necessarily lead to sparse estimate of differences. One solution to this problem is to use a "fused-lasso" regularizer $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_1\|_1$ to encourage the sparsity in changes between two parameters (Danaher et al., 2014).

However, applying sparsity inducing norms on individual models assumes that the true probabilistic models q_0 and q_1 have sparse parameters $\boldsymbol{\theta}_0^*$ and $\boldsymbol{\theta}_1^*$. In theoretical analysis, this leads to consistency results depending on the sparsity level of the individual model, the less sparse the individual models are, the worse the convergence rate is (see e.g., Theorem 1 in (Yang et al., 2012)). In reality, $\boldsymbol{\theta}_0^*$ and $\boldsymbol{\theta}_1^*$ may not be sparse, but the difference between them could be sparse.

To address this issue, previous works propose a density ratio-based approach to directly estimate the difference $\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_1^*$ (Liu et al., 2017; Kim et al., 2021). The intuition is that the ratio between exponential family models is

$$\frac{p(\boldsymbol{x};\boldsymbol{\theta}_1)}{p(\boldsymbol{x};\boldsymbol{\theta}_0)} \propto \exp\left(\langle\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0, f(\boldsymbol{x})\rangle\right),$$

determined entirely by the *differential parameter*. Thus,

fitting the density ratio function automatically learns the parameter change. Moreover, since this estimation is completely independent of individual parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, we do not need to regularize $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, eliminating the sparsity assumptions on $\boldsymbol{\theta}_0^*$ and $\boldsymbol{\theta}_1^*$.

7.2 Estimating Time-varying Probabilistic Models

While the density ratio approach learns the "discontinuous change" from $\boldsymbol{\theta}_0^*$ to $\boldsymbol{\theta}_1^*$, we may be interested in the *continuous process* from $\boldsymbol{\theta}_0^*$ to $\boldsymbol{\theta}_1^*$. We are interested in estimating $\boldsymbol{\theta}^*(t)$ given random samples $\boldsymbol{x}_t \sim q_t$. Naturally, one can fit model $p(\boldsymbol{x}; \boldsymbol{\theta}(t))$ to samples $\boldsymbol{x}_t$ at each time point. Assuming the time-varying process is "sparse", i.e., only a few parameters change with t, we can estimate $p(\boldsymbol{x}; \boldsymbol{\theta}(t))$ jointly using a multi-task learning objective with regularization $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(t')\|_1$ for a t' that is close to t. (Kolar and Xing, 2012; Hallac et al., 2017; Gibberd and Nelson, 2017). In this paper, we compare the proposed method with Loggle (Yang and Peng, 2020), which is designed to capture smoothly varying $\boldsymbol{\Theta}(t)$ and is a variant of the above algorithm.

However, if we are only interested in how $\boldsymbol{\theta}(t)$ changes with time t, i.e., the time-derivative $\partial_t \boldsymbol{\theta}(t)$, rather than the process $\boldsymbol{\theta}(t)$ itself. The aforementioned approach is again sub-optimal, as estimating a high dimensional, over-parameterized $\boldsymbol{\theta}(t)$ requires us to regularize the sparsity of $\boldsymbol{\theta}(t)$ for each t, thus again, putting unnecessary sparsity assumptions on $\boldsymbol{\theta}^*(t)$, leading to a consistency results depend on the sparsity level of each $\boldsymbol{\theta}^*(t)$ (see e.g., Theorem 1 in (Kolar and Xing, 2009).

8 Simulation Studies

We evaluate SparTSM and SparTSM+ estimator performance using datasets simulated with Gaussian Graphical Models (GGM). We sample $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}(t)^{-1})$ with $\boldsymbol{\Theta}(t) = \boldsymbol{\Theta}_0 + \boldsymbol{\Theta}'(t)$, where $\boldsymbol{\Theta}_0$ is a constant symmetric positive definite dense matrix, and $\boldsymbol{\Theta}'(t)$ is a *sparse* symmetric matrix that changes over time. Refer to Appendix G for settings of $\boldsymbol{\Theta}'(t)$ in each experiment.

8.1 Differential Parameters Estimation using SparTSM

We generate 5000 synthetic samples from a 20-node Gaussian Graphical Model (GGM) and conduct a performance comparison between Loggle and SparTSM. In this study, the non-zero elements of $\boldsymbol{\Theta}'(t)$ are formulated using a sine function (refer to Appendix G.1). In Figure 1, we illustrate the estimated $\partial_t \Theta_{i,j}(t)$ using SparTSM and the estimated $\Theta_{i,j}(t)$ with Loggle. The true time-variant parameters (where $\partial_t \Theta_{i,j}(t) \neq 0$) are marked in red. On one hand, SparTSM, which

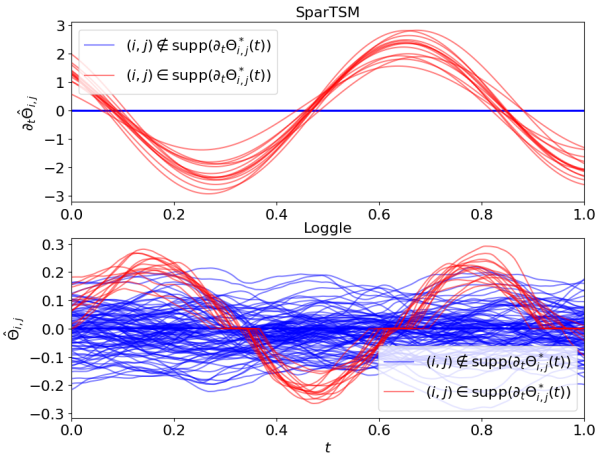
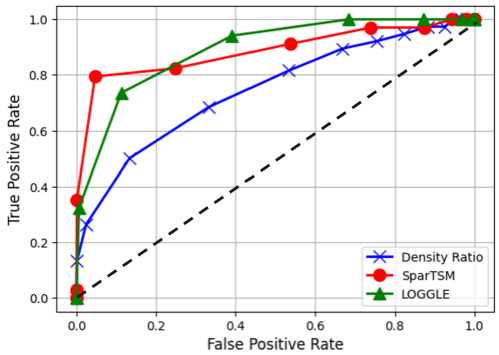
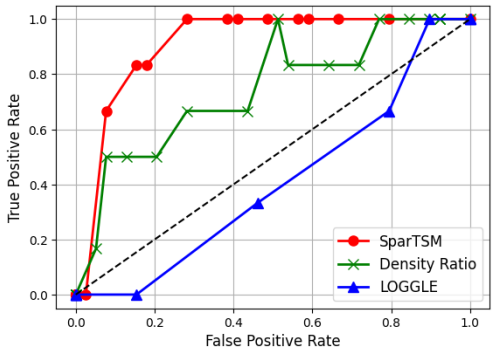


Figure 1: **SparTSM compared with Loggle.** Estimating the differential parameter vs. Estimating the time-varying parameter. The overall graphical model $\boldsymbol{\Theta}$ is non-sparse.

employs an appropriate basis function (Fourier basis function, see Section 4.4), succeeds in accurately estimating differential parameters, thereby clearly identifying the smooth transitions in the time-variant distribution. On the other hand, Loggle does identify time-varying parameters (red curves in Figure 1), but it does not immediately discern the changing parameters from the stationary parameters, as its estimates are conflated. This aligns with expectations, since our constructed $\boldsymbol{\Theta}(t)$ is consistently *non-sparse*, requiring the estimation of a *dense* parameter vector, which does not adhere to Loggle's sparsity assumptions. In the experiments below, we set $\phi(t) = t$. We quantitatively evaluate SparTSM's effectiveness in detecting parameter changes by comparing it to Loggle and the density ratio-based method (Liu et al., 2017; Kim et al., 2021), using ROC curves for assessment. We construct a GGM that changes linearly over time with 40 nodes and draw 1000 samples (Details can be found in G.1). The task of this experiment is to detect time-varying edges ($\partial_t \Theta_{i,j}(t) \neq 0$) in the GGM. For the density ratio approach, $\boldsymbol{\Theta}(1) - \boldsymbol{\Theta}(0)$ is calculated by estimating the ratio q_1/q_0, where we draw 500 samples from q_0 and 500 samples from q_1. By adjusting the regularization parameter in both SparTSM and the density ratio approach, we can modify the number of changes detected, resulting in an ROC curve. Note that Loggle does not directly provide $\partial_t \Theta_{i,j}(t)$, but rather offers a timeline of $\Theta_{i,j}(t)$ values. Since the true $\Theta_{i,j}(t)$ follows a linear trend, we estimate the time derivative using the least squares regression coefficients (see Appendix G.5). Different detection thresholds then yield a series of sensitivity levels. Note that this trick can only be applied when we know that the underlying change is



(a) **Time-varying edge detection on a Gaussian Graphical Model.** The diagonal line has been added for reference.



(b) **Time-varying edge detection on a truncated Gaussian Graphical Model.**

Figure 2: ROC curves of SparTSM, Density Ratio, and Loggle.

linear. The ROC curves are displayed in Figure 2(a), demonstrating that SparTSM's ROC curve achieves a performance comparable to Loggle and both SparTSM and Loggle significantly outperforms the density ratio method. We compute the average AUC value over 10 random trials in Tabel 1, and it can be seen that Loggle marginally outperforms SparTSM.

We restrict the domain of the Gaussian Graphical Model (GGM) to $\mathbb{R}_+^{10}$ and generate 2,000 samples from this truncated GGM (Lin et al., 2016b). The ROC curves of the three methods are shown in Figure 2(b), where the results demonstrate that both SparTSM and density ratio exhibit superior performance, while Loggle fails completely (AUC $\approx$ 0.5). This failure arises because the truncated GGM density contains an intractable normalizing constant, rendering the likelihood function in Loggle invalid. The density ratio and SparTSM method do not invovle normalizing constant calculation, and is not affected by the intractable normalising constant. The average AUC values in Table 1 further confirm this trend, with the SparTSM method significantly outperforming both density ratio and Loggle.

	SparTSM	Density Ratio	Loggle
GGM	0.875 (0.025)	0.738 (0.150)	**0.893 (0.026)**
Trunc. GGM	**0.733 (0.122)**	0.624 (0.153)	0.486 (0.120)

Table 1: Comparison of average AUCs over 10 trials.

Method	Deterministic	Random
Loggle	4.0%	6.0%
Oracle	3.4%	2.2%
SparTSM+	**5.6%**	**5.3%**

Table 2: The proportions of unsuccessful coverage at nominal confidence level of 95%. $\mathcal{H}_0 : \partial_t \Theta_{1,2}(t) = 0$.

8.2 Differential Parameter Inference using SparTSM+

In exploring SparTSM+, we assess it using two distinct linear GGM datasets, with details provided in Appendix G.2. We create 400 samples and 20 nodes from both fixed and random precision matrices, run SparTSM+ 1000 times, and plot the distribution of $(\tilde{\alpha}_{1,2} - \alpha^*_{1,2})/\hat{\sigma}_{1,2}$ in Figure 3. This showcases the effectiveness of the Gaussian approximation for the standardized SparTSM+. The histogram closely aligns with the standard normal density function. In the Q-Q plot, data points align with or are near the reference line, underscoring the precision of the Gaussian approximation and validating Theorem 6.5 and Corollary 6.7 through our experimental results.

We further compare SparTSM+ against the Oracle method and Loggle regarding confidence interval coverage across 1000 iterations. The Oracle approach presupposes known sparsity of elements and constructs confidence intervals using the asymptotic variance of an M-estimator (van der Vaart, 1998). Loggle's confidence intervals are derived from the 2.5% and 97.5% quantiles of the test statistic (estimated slope) from 100 permutation tests.

Table 2 shows the proportions of failure in achieving the nominal confidence level of 95%. Even with limited sample sizes, SparTSM+ maintains coverage close to the intended level, whereas the oracle is more conservative. Notably, our heuristic method for determining the confidence interval for Loggle also achieves fairly reliable coverage. Nonetheless, the permutation test is computationally intensive in practice.

As illustrated in Figure 4, the power of tests is demonstrated for $\partial_t \Theta_{1,2}(t)$ values between 0 and 10. The power is calculated as the ratio of rejections across 1000 independent trials at a significance level of 0.05

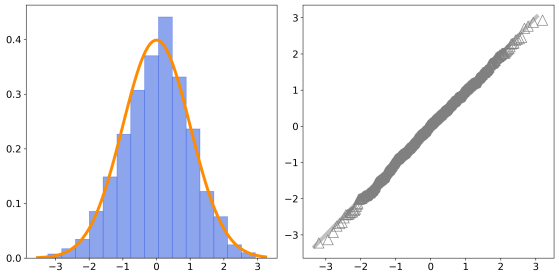


Figure 3: **Gaussian approximation** of SparTSM+. The left shows a bar plot and a QQ-plot is on the right.

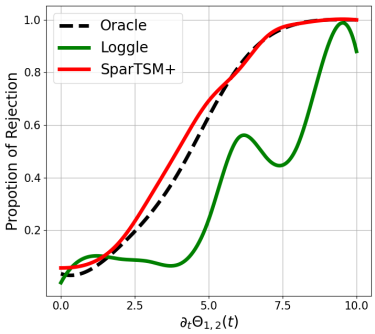


Figure 4: **Testing Power plot**. $\mathcal{H}_0 : \partial_t\Theta_{1,2}(t) = 0$.

under the null hypothesis $\mathcal{H}_0 : \partial_t\Theta_{1,2}(t) = 0$. It is evident that as $\partial_t\Theta_{1,2}(t)$ varies from $\mathcal{H}_0$, the rejection rate of our proposed method escalates, showcasing the test's efficacy. Notably, the proposed method usually provides greater power than the oracle, emphasizing its effectiveness in high-dimensional contexts.

9 Application: 109th US Senate

We employ SparTSM on the voting data from the 109th US Senate (Roy et al., 2016), which encapsulates the choices made by 100 senators over around two years. The data is organized as $\mathcal{D} = \{(t_i, \boldsymbol{x}_i)\}_{i=1}^{n=427}, \boldsymbol{x}_i \in \{0,1\}^{100}$, where '1' signifies a yea vote and '0' represents a nay vote. We consider these votes to adhere to a pairwise, time-dependent Ising model, described as $q_t \propto \exp\left(\sum_{i,j}\Theta_{i,j}(t)x_ix_j\right)$, and utilize SparTSM to estimate the differential parameter assuming a linear model $\Theta_{i,j}(t) = \alpha_{i,j}t$. The parameter λ_{lasso} is adjusted to ensure fewer than 100 non-zero elements remain in $\boldsymbol{\alpha}$. In Figure 5, we illustrate the differential graph $G = (V, E)$, where $E := \{(V_i, V_j)|\hat{\alpha}_{i,j} \neq 0\}$, meaning the edges denote variations in pairwise interactions within the Ising model. Nodes without any connection are excluded.

Notably, all calculated $\hat{\alpha}_{i,j}$ values are positive and occur within the same party. This suggests that as

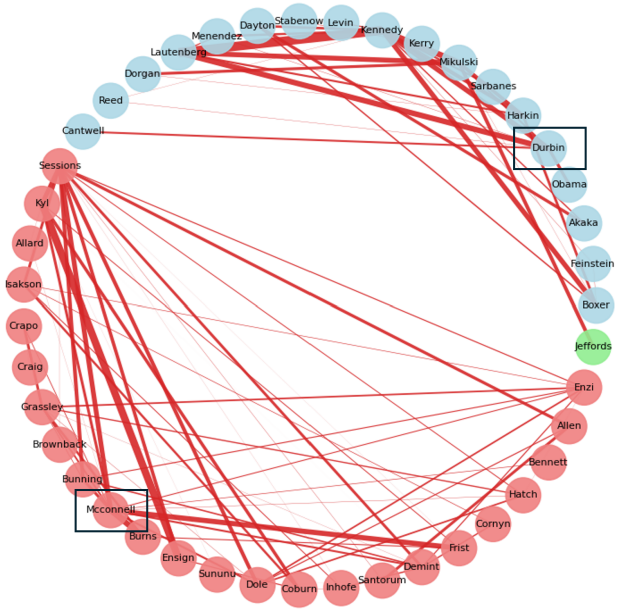


Figure 5: **The differential graph estimated from 109th senate voting dataset.** Red edge indicate a positive $\alpha_{i,j} \approx \partial_t\Theta^*_{i,j}(t)$. The widths of edges are proportional to $|\alpha_{i,j}|$. Democrats, republicans, independents are colored in blue, red and green respectively. Party whips are marked by a black rectangle.

the congressional term advances, senators increasingly align their votes with key party figures (like whips), creating "voting blocks" within the party. Furthermore, there is no apparent bipartisanship emerging between parties. In summary, these findings demonstrate a rise in partisanship throughout the congressional term.

10 Limitations and Future Works

Although the proposed method has strong theoretical properties and performs well in simulations, it has several limitations. First, it can only directly estimate the time differential parameter *within an exponential family*, as general probabilistic models cannot be parameterized solely using differential parameters. Second, the sufficient statistics $\boldsymbol{f}$ must be specified in advance, and selecting them for real-world datasets remains an open challenge. Third, the theoretical guarantees for the debiased estimator rely on the sparsity of the inverse Hessian of the objective function—an assumption also made in Kim et al. (2021) but not necessarily valid for all probabilistic models.

A promising future direction is to extend this work to settings where the sufficient statistics $\boldsymbol{f}$ can evolve over time, allowing them to be learned in a way that better captures the complexity of real-world data.

Acknowledgments

This work was carried out while Leyang Wang was on a University of Bristol School of Mathematics undergraduate summer bursary placement funded by the Heilbronn Institute for Mathematical Research. Daniel J. Williams was supported by the EPSRC Centre for Doctoral Training in Computational Statistics and Data Science, grant number EP/S023569/1. The research of Mladen Kolar is supported in part by NSF ECCS-2216912.

References

Barber, R. F. and Kolar, M. (2018). ROCKET: Robust confidence intervals via Kendall's tau for transelliptical graphical models. *The Annals of Statistics*, 46(6B):3422 – 3450.

Chen, L. H., Goldstein, L., and Shao, Q.-M. (2010). *Normal approximation by Stein's method*. Springer Science & Business Media.

Choi, K., Meng, C., Song, Y., and Ermon, S. (2022). Density ratio estimation via infinitesimal classification. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 2552–2573.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(2):373–397.

Drton, M. and Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4(1):365–393.

Gibberd, A. J. and Nelson, J. D. (2017). Regularized estimation of piecewise constant gaussian graphical models: The group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, 26(3):623–634.

Hallac, D., Park, Y., Boyd, S., and Leskovec, J. (2017). Network inference via the time-varying graphical lasso. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 205–213.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, Boca Raton.

Hoeffding, W. (1994). Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709.

Hyvärinen, A. (2007). Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512.

Kim, B., Liu, S., and Kolar, M. (2021). Two-Sample Inference for High-Dimensional Markov Networks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):939–962.

Kolar, M. and Xing, E. P. (2009). Sparsistent estimation of time-varying discrete markov random fields. *ArXiv e-prints, arXiv:0907.2337*.

Kolar, M. and Xing, E. P. (2011). On time varying undirected graphs. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 407–415.

Kolar, M. and Xing, E. P. (2012). Estimating networks with jumps. *Electronic Journal of Statistics*, 6(none):2069 – 2106.

Lin, L., Drton, M., and Shojaie, A. (2016a). Estimation of high-dimensional graphical models using regularized score matching. *Electronic journal of statistics*, 10(1):806.

Lin, L., Drton, M., and Shojaie, A. (2016b). Estimation of high-dimensional graphical models using regularized score matching. *Electronic journal of statistics*, 10(1):806.

Liu, S., Kanamori, T., and Williams, D. J. (2022). Estimating density models with truncation boundaries using score matching. *Journal of Machine Learning Research*, 23(186):1–38.

Liu, S., Suzuki, T., Relator, R., Sese, J., Sugiyama, M., and Fukumizu, K. (2017). Support consistency of direct sparse-change learning in Markov networks. *The Annals of Statistics*, 45(3):959 – 990.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. (2019). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363.

Lyu, S. (2009). Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, page 359–366.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2022). *Dataset shift in machine learning*. MIT Press.

Roy, S., Atchadé, Y., and Michailidis, G. (2016). Change Point Estimation in High Dimensional Markov Random-Field Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1187–1206.

Tang, L., Zhou, Y., Wang, L., Purkayastha, S., Zhang, L., He, J., Wang, F., and Song, P. X.-K. (2020). A review of multi-compartment infectious disease models. *International Statistical Review*, 88(2):462–513.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166 – 1202.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Williams, D. J. and Liu, S. (2022). Score matching for truncated density estimation on a manifold. In *Topological, Algebraic and Geometric Learning Workshops 2022*, pages 312–321. PMLR.

Xia, L., Nan, B., and Li, Y. (2023). Debiased lasso for generalized linear models with a diverging number of covariates. *Biometrics*, 79(1):344–357.

Yang, E., Allen, G., Liu, Z., and Ravikumar, P. (2012). Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, volume 25.

Yang, J. and Peng, J. (2020). Estimating time-varying graphical models. *Journal of Computational and Graphical Statistics*, 29(1):191–202.

Yu, M., Kolar, M., and Gupta, V. (2016). Statistical inference for pairwise graphical models using score matching. In *Advances in Neural Information Processing Systems*, volume 29.

Yu, S., Drton, M., and Shojaie, A. (2019). Generalized score matching for non-negative data. *Journal of Machine Learning Research*, 20(76):1–70.

Yu, S., Drton, M., and Shojaie, A. (2022). Generalized score matching for general domains. *Information and Inference: A Journal of the IMA*, 11(2):739–780.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):217–242.

Zhao, B., Wang, Y. S., and Kolar, M. (2019). Direct estimation of differential functional graphical models. *Advances in neural information processing systems*, 32.

Zhao, S., Cai, T., and Li, H. (2014). Direct estimation of differential networks. *Biometrika*, 101(2):253–268.

A Notations

We denote $\|\cdot\|$ as a norm on $\mathbb{R}^k$ and use $\|\boldsymbol{v}\|_l = \left(\sum_{i=1}^k |v_i|^l\right)^{1/l}$ as an usual ℓ_l norm for $l \in [1,\infty]$ and $\|\boldsymbol{v}\|_0 = |\mathrm{supp}(\boldsymbol{v})| = |\{i : v_i \neq 0\}|$; for a matrix $\boldsymbol{M} \in \mathbb{R}^{k\times k}$, $\|\boldsymbol{M}\| = \|\mathrm{Vec}(\boldsymbol{M})\|$; For $l > 0$, $|||\boldsymbol{M}|||_l = \sup_{\|v\|_0 \leq l, \|v\|=1} |v^\top \boldsymbol{M} v|$ is the maximum l-sparse eigenvalue of $\boldsymbol{M}$; consequently, $|||\boldsymbol{M}|||_\infty = \sup_{\|\boldsymbol{v}\|\leq 1} \|\boldsymbol{M}\boldsymbol{v}\|_\infty$.

B Proof of Theorem 3.1

Recall that

$$q(\boldsymbol{x};\boldsymbol{\theta}^*(t)) := \frac{\exp\left(\langle \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{x})\rangle\right)}{z(\boldsymbol{\theta}^*(t))}.$$

Thus,

$$\begin{aligned}
\partial_t \log q(\boldsymbol{x};\boldsymbol{\theta}^*(t)) &= \partial_t \langle \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{x})\rangle - \partial_t \log z(\boldsymbol{\theta}^*(t)) \\
&= \langle \partial_t \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{x})\rangle - \frac{\partial_t z(\boldsymbol{\theta}^*(t))}{z(\boldsymbol{\theta}^*(t))} \\
&= \langle \partial_t \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{x})\rangle - \frac{\partial_t \int_{\mathcal{X}} \exp\left(\langle \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{y})\rangle\right) d\boldsymbol{y}}{z(\boldsymbol{\theta}^*(t))} \\
&= \langle \partial_t \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{x})\rangle - \frac{\int_{\mathcal{X}} \partial_t \langle \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{y})\rangle \exp\left(\langle \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{y})\rangle\right) d\boldsymbol{y}}{z(\boldsymbol{\theta}^*(t))} \\
&= \langle \partial_t \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{x})\rangle - \frac{\int_{\mathcal{X}} \langle \partial_t \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{y})\rangle \exp\left(\langle \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{y})\rangle\right) d\boldsymbol{y}}{z(\boldsymbol{\theta}^*(t))} \\
&= \langle \partial_t \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{x})\rangle - \int_{\mathcal{X}} \langle \partial_t \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{y})\rangle \frac{\exp\left(\langle \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{y})\rangle\right)}{z(\boldsymbol{\theta}^*(t))} d\boldsymbol{y} \\
&= \langle \partial_t \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{x})\rangle - \int_{\mathcal{X}} \langle \partial_t \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{y})\rangle \, q(\boldsymbol{y};\boldsymbol{\theta}^*(t)) d\boldsymbol{y} \\
&= \langle \partial_t \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{x})\rangle - \mathbb{E}_{q_t}\left[\langle \partial_t \boldsymbol{\theta}^*(t), \boldsymbol{f}(\boldsymbol{y})\rangle\right],
\end{aligned}$$

as desired.

C Proof of Theorem 4.1

Let us begin from the initial formulation of our time based score matching objective with weight function $g(t) = g(t)$ for which $g(0) = g(1) = 0$, i.e. $g(t) = 0$ at the edges of our time domain $t \in [0,1]$. The initial objective is given by

$$\begin{aligned}
\mathcal{L}(\partial_t \boldsymbol{\theta}(t)) &= \int_{t=0}^{t=1} \mathbb{E}_{q_t}\left[g(t)\|\partial_t \log q_t(\boldsymbol{x}) - s_t\|^2\right] dt \\
&= \int_{t=0}^{t=1} \mathbb{E}_{q_t}\left[g(t)s_t^2\right] dt - 2\int_{t=0}^{t=1} \mathbb{E}_{q_t}\left[g(t)s_t \partial_t \log q_t(\boldsymbol{x})\right] dt + \int_{t=0}^{t=1} \mathbb{E}_{q_t}\left[g(t)\partial_t \log q_t(\boldsymbol{x})^2\right] dt \\
&= \int_{t=0}^{t=1} \mathbb{E}_{q_t}\left[g(t)s_t^2\right] dt - 2\int_{t=0}^{t=1} \int_{\mathcal{X}} \left[q_t \partial_t \log q_t(\boldsymbol{x})\right] g(t)s_t d\boldsymbol{x} dt + \int_{t=0}^{t=1} \mathbb{E}_{q_t}\left[g(t)\partial_t \log q_t(\boldsymbol{x})^2\right] dt \\
&= \int_{t=0}^{t=1} \mathbb{E}_{q_t}\left[g(t)s_t^2\right] dt - 2\int_{t=0}^{t=1} \int_{\mathcal{X}} \partial_t q_t g(t)s_t d\boldsymbol{x} dt + \int_{t=0}^{t=1} \mathbb{E}_{q_t}\left[g(t)\partial_t \log q_t(\boldsymbol{x})^2\right] dt,
\end{aligned}$$

where in the final line we have used $q_t \partial_t \log q_t(\boldsymbol{x}) = q_t \partial_t \log q_t = \partial_t q_t$ to simplify. First note that the final term is a constant with respect to the model s_t, and so we can write $C = \int_{t=0}^{t=1} \mathbb{E}_{q_t}\left[g(t)\partial_t \log q_t(\boldsymbol{x})^2\right] dt$. We continue by expanding the middle term via integration by parts

$$\mathcal{L}(\partial_t \boldsymbol{\theta}(t)) = \int_{t=0}^{t=1} \mathbb{E}_{q_t}\left[g(t)s_t^2\right] dt - 2\left[\int_{\mathcal{X}} q_t g(t)s_t d\boldsymbol{x}\right]_{t=0}^{t=1} + 2\int_{t=0}^{t=1} \int_{\mathcal{X}} q_t \partial_t (g(t)s_t) d\boldsymbol{x} dt + C$$

$$
\begin{aligned}
&= \int_{t=0}^{t=1} \mathbb{E}_{q_t}\left[g(t)s_t^2\right] dt + 2\int_{t=0}^{t=1}\int_{\mathcal{X}} q_t \partial_t(g(t)s_t) d\boldsymbol{x}dt + C \\
&= \int_{t=0}^{t=1} g(t)\mathbb{E}_{q_t}\left[s_t^2\right] dt + 2\int_{t=0}^{t=1} \partial_t g(t)\mathbb{E}_{q_t}\left[s_t\right] dt + 2\int_{t=0}^{t=1} g(t)\mathbb{E}_{q_t}\left[\partial_t s_t\right] dt + C, \qquad (21)
\end{aligned}
$$

where the second equality is due to $\left[\int_{\mathcal{X}} q_t g(t) s_t d\boldsymbol{x}\right]_{t=0}^{t=1} = 0$ as $g(0) = g(1) = 0$.

Let us now substitute the model given in Equation (4), stated again here for completeness, given by

$$
s_t = s(\boldsymbol{x}, t) = \langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x})\rangle - \mathbb{E}_{q_t}\left[\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y})\rangle\right].
$$

We calculate $\mathbb{E}_{q_t}[s_t]$ and $\mathbb{E}_{q_t}[\partial_t s_t]$ to substitute into the equation above. These are given by

$$
\begin{aligned}
\mathbb{E}_{q_t}[s_t] &= \mathbb{E}_{q_t}\left[\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x})\rangle - \mathbb{E}_{q_t}\left[\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y})\rangle\right]\right] \\
&= \mathbb{E}_{q_t}\left[\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x})\rangle\right] - \mathbb{E}_{q_t}\left[\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y})\rangle\right] \\
&= 0 \\
\mathbb{E}_{q_t}[\partial_t s_t] &= \mathbb{E}_{q_t}\left[\partial_t\left(\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x})\rangle - \mathbb{E}_{q_t}\left[\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y})\rangle\right]\right)\right] \\
&= \mathbb{E}_{q_t}\left[\left\langle \partial_t^2 \boldsymbol{\theta}_t, \boldsymbol{f}(\boldsymbol{x})\right\rangle - \partial_t\left(\mathbb{E}_{q_t}\left[\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y})\rangle\right]\right)\right] \\
&= \mathbb{E}_{q_t}\left[\left\langle \partial_t^2 \boldsymbol{\theta}_t, \boldsymbol{f}(\boldsymbol{x})\right\rangle - \partial_t\left(\int_{\mathcal{X}} q_t \left\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y})\right\rangle d\boldsymbol{y}\right)\right] \\
&= \mathbb{E}_{q_t}\left[\left\langle \partial_t^2 \boldsymbol{\theta}_t, \boldsymbol{f}(\boldsymbol{x})\right\rangle - \int_{\mathcal{X}} \partial_t q_t \left\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y})\right\rangle d\boldsymbol{y} - \int_{\mathcal{X}} q_t \left\langle \partial_t^2 \boldsymbol{\theta}_t, \boldsymbol{f}(\boldsymbol{y})\right\rangle d\boldsymbol{y}\right] \\
&= \mathbb{E}_{q_t}\left[\left\langle \partial_t^2 \boldsymbol{\theta}_t, \boldsymbol{f}(\boldsymbol{x})\right\rangle - \int_{\mathcal{X}} \partial_t q_t \left\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y})\right\rangle d\boldsymbol{y} - \mathbb{E}_{q_t}\left\langle \partial_t^2 \boldsymbol{\theta}_t, \boldsymbol{f}(\boldsymbol{y})\right\rangle\right] \\
&= \mathbb{E}_{q_t}\left[\left\langle \partial_t^2 \boldsymbol{\theta}_t, \boldsymbol{f}(\boldsymbol{x})\right\rangle - \int_{\mathcal{X}} \partial_t q_t \left\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y})\right\rangle d\boldsymbol{y}\right] - \mathbb{E}_{q_t}\left[\left\langle \partial_t^2 \boldsymbol{\theta}_t, \boldsymbol{f}(\boldsymbol{y})\right\rangle\right] \\
&= \mathbb{E}_{q_t}\left[-\int_{\mathcal{X}} \partial_t q_t \left\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y})\right\rangle d\boldsymbol{y}\right],
\end{aligned}
$$

where the integral $\int_{\mathcal{X}}(\dots)d\boldsymbol{y}$ is the same integral as $\int_{\mathcal{X}}(\dots)d\boldsymbol{x}$ but written as such to make them distinct from one another. Substituting these into Equation (21) gives

$$
\begin{aligned}
\mathcal{L}(\partial_t \boldsymbol{\theta}(t)) &= \int_{t=0}^{t=1} g(t)\mathbb{E}_{q_t}\left[s_t^2\right] dt - 2\int_{t=0}^{t=1} g(t)\mathbb{E}_{q_t}\left[\int_{\mathcal{X}} \partial_t q_t \left\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y})\right\rangle d\boldsymbol{y}\right] dt + C \\
&= \int_{t=0}^{t=1} g(t)\mathbb{E}_{q_t}\left[s_t^2\right] dt - 2\int_{t=0}^{t=1} g(t)\int_{\mathcal{X}} q_t\left(\int_{\mathcal{X}} \partial_t q_t \left\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y})\right\rangle d\boldsymbol{y}\right) d\boldsymbol{x}dt + C \\
&\overset{(a)}{=} \int_{t=0}^{t=1} g(t)\mathbb{E}_{q_t}\left[s_t^2\right] dt - 2\int_{t=0}^{t=1} g(t)\left(\int_{\mathcal{X}} q_t d\boldsymbol{x}\right)\int_{\mathcal{X}} \partial_t q_t \left\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{y})\right\rangle d\boldsymbol{y}dt + C \\
&\overset{(b)}{=} \int_{t=0}^{t=1} g(t)\mathbb{E}_{q_t}\left[s_t^2\right] dt - 2\int_{t=0}^{t=1} g(t)\int_{\mathcal{X}} \partial_t q_t \left\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x})\right\rangle d\boldsymbol{x}dt + C.
\end{aligned}
$$

In the equality denoted by (a), we have used the fact that inside the integral $\int_{\mathcal{X}} q_t(\dots)d\boldsymbol{x}$ the only variable dependent on $\boldsymbol{x}$ was q_t, as $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent. We also have that $\int_{\mathcal{X}} q_t d\boldsymbol{x} = 1$ by q_t being a probability density function. The equality denoted by (b) contains a re-labelling of $\boldsymbol{x} = \boldsymbol{y}$, as they are the same variable, only labelled differently originally to make them distinct. We use integration by parts one final time to obtain

$$
\begin{aligned}
\mathcal{L}(\partial_t \boldsymbol{\theta}(t)) &= \int_{t=0}^{t=1} g(t)\mathbb{E}_{q_t}\left[s_t^2\right] dt - 2\left[\int_{\mathcal{X}} q_t g(t) \left\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x})\right\rangle d\boldsymbol{x}\right]_{t=0}^{t=1} + 2\int_{t=0}^{t=1}\int_{\mathcal{X}} q_t \partial_t\left(g(t)\left\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x})\right\rangle\right) d\boldsymbol{x}dt + C \\
&= \int_{t=0}^{t=1} g(t)\mathbb{E}_{q_t}\left[s_t^2\right] dt + 2\int_{t=0}^{t=1}\int_{\mathcal{X}} q_t \partial_t\left(g(t)\left\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x})\right\rangle\right) d\boldsymbol{x}dt + C,
\end{aligned}
$$

using again that $g(0) = g(1) = 0$, and finally

$$
\mathcal{L}(\partial_t \boldsymbol{\theta}(t)) = \int_{t=0}^{t=1} g(t)\mathbb{E}_{q_t}\left[s_t^2\right] dt + 2\int_{t=0}^{t=1}\int_{\mathcal{X}} q_t \partial_t g(t)\left\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x})\right\rangle d\boldsymbol{x}dt + 2\int_{t=0}^{t=1}\int_{\mathcal{X}} q_t g(t)\left\langle \partial_t^2 \boldsymbol{\theta}_t, \boldsymbol{f}(\boldsymbol{x})\right\rangle d\boldsymbol{x}dt + C
$$

$$= \int_{t=0}^{t=1} \mathbb{E}_{q_t} \left[g(t) s_t^2 + 2\partial_t g(t) \left\langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(\boldsymbol{x}) \right\rangle + 2g(t) \left\langle \partial_t^2 \boldsymbol{\theta}_t, \boldsymbol{f}(\boldsymbol{x}) \right\rangle \right] dt$$

which is the same as Equation (6) in the main text.

D Finite-sample Estimation Error of Lasso Estimator

An equivalent sample objective function is as the following:

$$\hat{\mathcal{L}}(\boldsymbol{\alpha}) = \frac{1}{2n} \left[\boldsymbol{\alpha} \tilde{\boldsymbol{F}}^\top \boldsymbol{G} \tilde{\boldsymbol{F}} \boldsymbol{\alpha}^\top + 2 \mathbf{1}_n^\top \partial_t \boldsymbol{G} \boldsymbol{F} \boldsymbol{\alpha}^\top \right] \tag{22}$$

where diagonal matrix $\boldsymbol{G}, \partial_t \boldsymbol{G} \in \mathbb{R}^{n \times n}$ with i-th diagonal entry to be $g(t_i)$ and $\partial_t g(t_i)$ respectively.

Theorem D.1. *Suppose Assumption 6.1 and 6.3 hold. Any minimizer of the objective function Equation (12) with regularization parameter lower bounded as* $\lambda_{\text{lasso}} \geq 2 \left\| \nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*) \right\|_\infty$ *satisfies*

$$||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*||_2 \leq \frac{3}{\kappa} \|\boldsymbol{\alpha}^*\|_0^{1/2} \lambda_{\text{lasso}}, \quad ||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*||_1 \leq \frac{6}{\kappa} \|\boldsymbol{\alpha}^*\|_0 \lambda_{\text{lasso}}. \tag{23}$$

D.1 Proof of Theorem D.1

Following equation 22 and Assumption 6.1, we can express the second order Taylor polynomial around $\boldsymbol{\alpha}^*$ as follows:

$$\hat{\mathcal{L}}(\hat{\boldsymbol{\alpha}}) = \hat{\mathcal{L}}(\boldsymbol{\alpha}^*) + \frac{1}{n} \left[\boldsymbol{\alpha}^* \tilde{\boldsymbol{F}}^\top \boldsymbol{G} \tilde{\boldsymbol{F}} + \mathbf{1}_n^\top \partial_t \boldsymbol{G} \boldsymbol{F} \right] (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)^\top + \frac{1}{2n} (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \left[\tilde{\boldsymbol{F}}^\top \boldsymbol{G} \tilde{\boldsymbol{F}} \right] (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)^\top \tag{24}$$

where the residual $R_2(\hat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}^*) = 0$ since $\hat{\mathcal{L}}$ quadratic.

Lemma D.2. *Under condition* $\lambda_{\text{lasso}} \geq 2 \left\| \nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*) \right\|_\infty$*, the error vector* $\widehat{\Delta} = \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^* \in \mathbb{C}_3(S)$

Proof. Since $\hat{\boldsymbol{\alpha}}$ is optimal, we have

$$\hat{\mathcal{L}}(\hat{\boldsymbol{\alpha}}) + \lambda_{\text{lasso}} ||\hat{\boldsymbol{\alpha}}||_1 \leq \hat{\mathcal{L}}(\boldsymbol{\alpha}^*) + \lambda_{\text{lasso}} ||\boldsymbol{\alpha}^*||_1 \tag{25}$$

Rearranging we have from second order Taylor approximation:

$$0 \leq \frac{1}{2n} ||\boldsymbol{G}^{\frac{1}{2}} \tilde{\boldsymbol{F}} \hat{\Delta}^\top||_2^2 \leq \frac{1}{n} |\nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*) \hat{\Delta}^\top| + \lambda_{\text{lasso}} \{ ||\boldsymbol{\alpha}^*||_1 - ||\hat{\boldsymbol{\alpha}}||_1 \} \tag{26}$$

where $|\cdot|$ denote the absolute value. Now since $\boldsymbol{\alpha}^*$ is S-sparse, we can write

$$\|\boldsymbol{\alpha}^*\|_1 - \|\hat{\boldsymbol{\alpha}}\|_1 = \|\boldsymbol{\alpha}_S^*\|_1 - \|\boldsymbol{\alpha}_S^* + \widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1. \tag{27}$$

Using Holder's inequality and the triangle inequality, we have

$$\begin{aligned} 0 \leq \frac{1}{n} ||G^{\frac{1}{2}} \tilde{\boldsymbol{F}} \hat{\Delta}^\top||_2^2 &\leq 2 |\nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*) \hat{\Delta}^\top| + 2\lambda_{\text{lasso}} \{ \|\boldsymbol{\alpha}_S^*\|_1 - \|\boldsymbol{\alpha}_S^* + \widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1 \} && (28) \\ &\leq 2 \|\nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\|_\infty \|\widehat{\Delta}\|_1 + 2\lambda_{\text{lasso}} \{ \|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1 \} && (29) \\ &\leq \lambda_{\text{lasso}} \{ 3 \|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1 \}, && (30) \end{aligned}$$

where last inequality shows that $\widehat{\Delta} \in \mathbb{C}_3(S)$. □

Lemma D.3. $||\widehat{\Delta}||_1 \leq 2\sqrt{s} ||\widehat{\Delta}||_2$ *where* $s = |S|$.

Proof. Since S is the support of $\boldsymbol{\alpha}^*$

$$||\boldsymbol{\alpha}_S^*||_1 = ||\boldsymbol{\alpha}^*||_1 \geq ||\boldsymbol{\alpha}^* + \widehat{\Delta}||_1 = ||\boldsymbol{\alpha}_S^* + \widehat{\Delta}_S||_1 + ||\widehat{\Delta}_{S^c}||_1 \geq ||\boldsymbol{\alpha}_S^*||_1 - ||\widehat{\Delta}_S||_1 + ||\widehat{\Delta}_{S^c}||_1. \tag{31}$$

where we used the fact that $\boldsymbol{\alpha}_{S^c}^* = 0$ and triangle inequality. This implies that $\widehat{\Delta} \in \mathbb{C}_1(S)$. Therefore

$$||\widehat{\Delta}||_1 = ||\widehat{\Delta}_S||_1 + ||\widehat{\Delta}_{S^c}||_1 \leq 2||\widehat{\Delta}_S||_1 \leq 2\sqrt{s} ||\widehat{\Delta}||_2 \tag{32}$$

□

With all above, we can then apply the RE condition. Finally we have $||\widehat{\Delta}||_2^2 \leq \frac{3}{\kappa}\lambda_{\text{lasso}}\sqrt{s}||\widehat{\Delta}||_2$, which implies that

$$||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*||_2 \leq \frac{3}{\kappa}\lambda_{\text{lasso}}\sqrt{s} \tag{33}$$

D.2 Proof of Theorem 6.4

We have from definition that $\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{j=1}^{n}\nabla_{\boldsymbol{\alpha}}\boldsymbol{m}_{\boldsymbol{\alpha}}(t_j, \boldsymbol{x}_j)$.

Lemma D.4. *Under the condition that the r.v. $\boldsymbol{f}(x)$ is bounded in ℓ_∞ norm, let $\sigma^2 = \max_{1\leq i\leq k}\boldsymbol{\Sigma}_{ii}$, where $\boldsymbol{\Sigma}$ is the covariance matrix of the random variable $\nabla_{\boldsymbol{\alpha}}\boldsymbol{m}_{\boldsymbol{\alpha}^*}(t, \boldsymbol{x})$, the elements of $\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}(\boldsymbol{\alpha}^*)$ follow a zero-mean sub-Gaussian distribution with parameter σ^2/n.*

Proof. **(1) Proof the sub-Gaussian:** let $|g(t)| \leq C_g \in \mathbb{R}_+, |\partial_t g(t)| \leq C_{\partial_t g(t)} \in \mathbb{R}_+, ||\mathbb{E}_{q_t}[\boldsymbol{f}(x)]||_\infty \leq C_E \in \mathbb{R}_+$ and $||\boldsymbol{f}(x)||_\infty \leq C_f \in \mathbb{R}_+$ for $t \in [0, 1]$. By using triangle inequality in $(i), (iii)$ and Cauchy-Schwarz inequality in (ii) , we have by definition of $\nabla_{\boldsymbol{\alpha}}\boldsymbol{m}_{\boldsymbol{\alpha}}(t, \boldsymbol{x})$

$$\|\nabla_{\boldsymbol{\alpha}}\boldsymbol{m}_{\boldsymbol{\alpha}}(t, \boldsymbol{x})\|_\infty \leq \|\boldsymbol{\alpha}^*\|_\infty \left(C_g(C_f + C_E)^2 + C_f C_{\partial_t g(t)}\right) \tag{34}$$

therefore by $M = ||\boldsymbol{\alpha}^*||(C_g(C_f + C_E)^2 + C_f C_{\partial_t g(t)})$ for simplicity

$$\left\|\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}\left(\boldsymbol{\alpha}^*\right)\right\|_\infty = \left\|\frac{1}{n}\sum_{j=1}^{n}\nabla_{\boldsymbol{\alpha}}\boldsymbol{m}_{\boldsymbol{\alpha}}(t_j, \boldsymbol{x}_j)\right\|_\infty \leq M \tag{35}$$

which implies that $\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}(\boldsymbol{\alpha}^*)$ is element-wise bounded hence all elements are sub-Gaussian by (Hoeffding, 1994). Therefore, fixed number of data points n, then each element of $\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}(\boldsymbol{\alpha}^*)$ follows a sub-Gaussian distribution with parameter σ^2/n by addition rule of variance.

(2) Proof of zero-mean: We also assert that $\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}(\boldsymbol{\alpha}^*) \in \mathbb{R}^k$ is a zero-mean random variable. Recall that the conditional density function $q(x|t)$ as $q_t(x)$ where $q_t : \mathbb{R}^k \to \mathbb{R}$ and $q : \mathbb{R} \to \mathbb{R}$ represent the density of t, where t is uniformly distributed in the domain $[0, 1]$, we then have the following:

$$\begin{aligned}
&\mathbb{E}_{x\sim q_t(x),\, t\sim q(t)}\left[\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\right] \\
&= \frac{1}{n}\mathbb{E}_{x\sim q_t(x),\, t\sim q(t)}\left[\boldsymbol{\alpha}^*\tilde{\boldsymbol{F}}^\top \boldsymbol{G}\tilde{\boldsymbol{F}} + \mathbf{1}_n^\top \partial_t \boldsymbol{G}\boldsymbol{F}\right] && (36)\\
&\overset{(i)}{=} \mathbb{E}_{x\sim q_t(x),\, t\sim q(t)}\left[\boldsymbol{\alpha}^*(\boldsymbol{f}(x) - \mathbb{E}_{q_t}[\boldsymbol{f}(y)])g(t)(\boldsymbol{f}(x) - \mathbb{E}_{q_t}[\boldsymbol{f}(y)]) + (\partial_t g(t))\boldsymbol{f}(x)\right] && (37)\\
&= \int_t\int_x q(t)q_t(x)\{\boldsymbol{\alpha}^*(\boldsymbol{f}(x) - \mathbb{E}_{q_t}[\boldsymbol{f}(y)])g(t)\left(\boldsymbol{f}(x) - \mathbb{E}_{q_t}[\boldsymbol{f}(y)]\right) + (\partial_t g(t))\boldsymbol{f}(x)\}\mathrm{d}x\,\mathrm{d}t && (38)\\
&\overset{(ii)}{=} \int_t\int_x q_t(x)\{(\partial_t \log q_t(x))g(t)\left(\boldsymbol{f}(x) - \mathbb{E}_{q_t}[\boldsymbol{f}(y)]\right) + (\partial_t g(t))\boldsymbol{f}(x)\}\mathrm{d}x\,\mathrm{d}t && (39)\\
&= \int_t\int_x (\partial_t q_t(x))g(t)\left(\boldsymbol{f}(x) - \mathbb{E}_{q_t}[\boldsymbol{f}(y)]\right) + q_t(x)(\partial_t g(t))\boldsymbol{f}(x)\}\mathrm{d}x\,\mathrm{d}t && (40)\\
&= \left[q_t(x)g(t)\left(\boldsymbol{f}(x) - \mathbb{E}_{q_t}[\boldsymbol{f}(y)]\right)\right]_{t=0}^{t=1} - \int_t\int_x q_t(x)\partial_t\left[g(t)\left(\boldsymbol{f}(x) - \mathbb{E}_{q_t}[\boldsymbol{f}(y)]\right)\right] - q_t(x)(\partial_t g(t))\boldsymbol{f}(x)\mathrm{d}x\,\mathrm{d}t && (41)\\
&\overset{(iii)}{=} \int_t\int_x -q_t(x)\left[(\partial_t g(t))\left(\boldsymbol{f}(x) - \mathbb{E}_{q_t}[\boldsymbol{f}(y)]\right) - g(t)\partial_t\mathbb{E}_{q_t}\left[\boldsymbol{f}(y)\right]\right] + q_t(x)(\partial_t g(t))\boldsymbol{f}(x)\mathrm{d}x\,\mathrm{d}t && (42)\\
&= \int_t\int_x q_t(x)\left([\partial_t g(t)]\mathbb{E}_{q_t}[\boldsymbol{f}(y)] + g(t)\partial_t\mathbb{E}_{q_t}[\boldsymbol{f}(y)]\right)\mathrm{d}x\,\mathrm{d}t = \int_t\int_x q_t(x)\partial_t[g(t)\mathbb{E}_{q_t}[\boldsymbol{f}(y)]]\mathrm{d}x\,\mathrm{d}t && (43)\\
&= [q_t(x)g(t)\mathbb{E}_{q_t}[\boldsymbol{f}(y)]]_{t=0}^{t=1} - \int_t\int_x (\partial_t q_t(x))g(t)\mathbb{E}_{q_t}[\boldsymbol{f}(y)]\mathrm{d}x\,\mathrm{d}t && (44)\\
&\overset{(iv)}{=} -\int_t \partial_t\left(\int_x q_t(x)\mathrm{d}x\right)g(t)\mathbb{E}_{q_t}[\boldsymbol{f}(y)]\mathrm{d}t = \mathbf{0} && (45)
\end{aligned}$$

where (i) follows that $\{f(x_1), .., f(x_n)\}$ are i.i.d; (ii) follows from the definition of $\partial_t \log q_t(x)$ and $q(t) = 1$; $(iii), (iv)$ uses the boundary condition of $g(t)$. Therefore, each element of $\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}(\boldsymbol{\alpha}^*)$ is a zero-mean sub-Gaussian r.v.

□

Lemma D.5. *Let J_i denotes i-th element of $\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}(\boldsymbol{\alpha}^*)$, we have*

$$\mathbb{E}[\max_{i\in[k]}|J_i|] \leq 2\sqrt{\frac{\sigma^2\log k}{n}} \tag{46}$$

Proof. We have $\{J_i\}_{i=1}^k$ is a sequence of zero-mean random variables, each follows sub-Gaussian with parameter σ^2/n. For any $\lambda > 0$, we can use the convexity of the exponential function to obtain

$$\exp\{\lambda\mathbb{E}[\max_{i\in[k]} J_i]\} \leq \mathbb{E}[\exp\{\lambda \max_{i\in[k]} J_i\}]$$

by Jensen's inequality. And by the monotonicity of the exponential

$$\mathbb{E}[\exp\{\lambda \max_{i\in[k]} J_i\}] = \mathbb{E}[\max_{i\in[k]} e^{\lambda J_i}] \leq \sum_{i=1}^{k}\mathbb{E}[e^{\lambda J_i}] \leq ke^{\frac{\lambda^2\sigma^2}{2n}}.$$

where last inequality follows the definition of sub-Gaussian r.v. Therefore,

$$\mathbb{E}[\max_{i\in[k]} J_i] \leq \frac{\log k}{\lambda} + \lambda\frac{\sigma^2}{2n}$$

and $\lambda = \sqrt{\frac{2n\log k}{\sigma^2}}$ is optimal in $\lambda > 0$; substituting we have

$$\mathbb{E}\left[\max_{i\in[k]} J_i\right] \leq \frac{\sigma}{\sqrt{2n}}\sqrt{\log k} + \frac{\sigma}{\sqrt{2n}}\sqrt{\log k} = \sqrt{2\sigma^2\log k/n} \tag{47}$$

Since 47 does not assume independence between individual J_i. The result follows by

$$\max_{i\in[k]}|J_i| = \max\{|J_1|,\ldots,|J_k|\} = \max\{J_1,\ldots,J_k,-J_1,\ldots,-J_k\}$$

□

Consequently, from standard sub-Gaussian tail bound and definition of ℓ_∞ norm, we have

$$\mathbb{P}\left[\left\|\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\right\|_\infty \geq \sigma\left(\sqrt{\frac{2\log k}{n}} + \delta\right)\right] \leq 2e^{-\frac{n\delta^2}{2}} \text{ for all } \delta > 0 \tag{48}$$

Hence if we set $\lambda_{\text{lasso}} = 2\sigma\left(\sqrt{\frac{2\log k}{n}} + \delta\right)$, then we have the probability that $\lambda_{\text{lasso}} \geq 2||\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}(\boldsymbol{\alpha}^*)||_\infty$ is in rate $1 - 2\exp\{-\Theta(n\delta^2)\}$, which implies

$$||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*||_2 \leq \frac{6}{\kappa}\sqrt{s}\sigma\left(\sqrt{\frac{2\log k}{n}} + \delta\right) \tag{49}$$

with probability greater than $1 - 2\exp\{-\frac{n\delta^2}{2}\}$ for all $\delta > 0$.

E Theoretical Results of Debiased lasso

E.1 Variance Estimator

Lemma E.1. $|||\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}) - \hat{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}^*)|||_\infty \leq L||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*||_1$ *where L is a constant.*

Proof. Apply the fact that there exists L_0 such that $\|\nabla_{\boldsymbol{\alpha}} m(\hat{\boldsymbol{\alpha}}) - \nabla_{\boldsymbol{\alpha}} m(\boldsymbol{\alpha}^*)\| \leq L_0\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|$ after computing the form of each $\hat{\boldsymbol{\Sigma}}_{k,k'}(\hat{\boldsymbol{\alpha}}) - \hat{\boldsymbol{\Sigma}}_{k,k'}(\boldsymbol{\alpha}^*)$. □

Lemma E.2. *On the event that*

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 \le \delta_\alpha/2, \quad \left\|\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right\|_1 \le \delta_{\boldsymbol{\omega}}, \quad \textit{and} \quad |||\hat{\boldsymbol{\Sigma}}(\alpha^*) - \boldsymbol{\Sigma}|||_\infty \le \delta_\sigma/2,$$

we have

$$\left|\hat{\sigma}_j^2 - \sigma_j^2\right| \le (L\delta_\alpha + \delta_\sigma)\left(\left\|\boldsymbol{\omega}_j^*\right\|_1^2 + \delta_{\boldsymbol{\omega}}^2\right) + \delta_{\boldsymbol{\omega}}^2 |||\boldsymbol{\Sigma}|||_\infty \tag{50}$$

Proof. We have by definition of $\hat{\sigma}_j^2$

$$\begin{aligned}
&\hat{\sigma}_j^2 - \sigma_j^2 = \tilde{\boldsymbol{\omega}}_j^\top \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}})\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^{*\top}\Sigma\boldsymbol{\omega}_j^* && (51)\\
&\implies \left|\tilde{\boldsymbol{\omega}}_j^\top \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}})\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^{*\top}\Sigma\boldsymbol{\omega}_j^*\right| && (52)\\
&\le \left|\tilde{\boldsymbol{\omega}}_j^\top \left(\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}) - \Sigma\right)\tilde{\boldsymbol{\omega}}_j\right| + \left|(\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*)^\top \Sigma(\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*)\right| && (53)\\
&\le |||\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}) - \Sigma|||_\infty \left\|\tilde{\boldsymbol{\omega}}_j\right\|_1^2 + |||\Sigma|||_\infty \left\|\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right\|_1^2 && (54)\\
&\le \left(|||\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}) - \hat{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}^*)|||_\infty + |||\hat{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}^*) - \Sigma|||_\infty\right)\left\|\tilde{\boldsymbol{\omega}}_j\right\|_1^2 + |||\boldsymbol{\Sigma}|||_\infty \left\|\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right\|_1^2 && (55)\\
&\le \left(L\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + |||\hat{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}^*) - \boldsymbol{\Sigma}|||_\infty\right)\left\|\tilde{\boldsymbol{\omega}}_j\right\|_1^2 + |||\boldsymbol{\Sigma}|||_\infty \left\|\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right\|_1^2 && (56)\\
&\le (L\delta_{\boldsymbol{\alpha}} + \delta_\sigma)\left(\left\|\boldsymbol{\omega}_j^*\right\|_1^2 + \delta_{\boldsymbol{\omega}}^2\right) + |||\boldsymbol{\Sigma}|||_\infty \delta_{\boldsymbol{\omega}}^2 && (57)
\end{aligned}$$

as desired. □

Lemma E.3. *There exists constants c_0, c, c' depend only on M such that for any $t \in [c_0\sqrt{\frac{\log k}{n}}, 1]$ such that*

$$\mathbb{P}\left(\|\hat{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}^*) - \boldsymbol{\Sigma}\|_\infty \ge t\right) \le c\exp\{-c't^2 n\} \tag{58}$$

Proof. We have by denoting sample mean as $\hat{\boldsymbol{\mu}}$ and true mean as $\boldsymbol{\mu}$, we have for any $j, j' \in [k]$

$$\begin{aligned}
&\hat{\boldsymbol{\Sigma}}_{j,j'}(\boldsymbol{\alpha}^*) - \boldsymbol{\Sigma}_{j,j'} = && (59)\\
&\left(\frac{1}{n}\sum_{i=1}^n \left([\nabla_{\boldsymbol{\alpha}} \boldsymbol{m}_{\boldsymbol{\alpha}}(\boldsymbol{x}_i, t_i)]_j - \boldsymbol{\mu}_j\right)\left([\nabla_{\boldsymbol{\alpha}} \boldsymbol{m}_{\boldsymbol{\alpha}}(\boldsymbol{x}_i, t_i)]_{j'} - \boldsymbol{\mu}_{j'}\right) - \boldsymbol{\Sigma}_{j,j'}\right) - \left(\boldsymbol{\mu}_j - \hat{\boldsymbol{\mu}}_j\right)\left(\boldsymbol{\mu}_{j'} - \hat{\boldsymbol{\mu}}_{j'}\right) && (60)
\end{aligned}$$

Suppose t satisfy the condition stated in lemma, and suppose

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n \left([\nabla_{\boldsymbol{\alpha}} \boldsymbol{m}_{\boldsymbol{\alpha}}(\boldsymbol{x}_i, t_i)]_j - \boldsymbol{\mu}_j\right)\left([\nabla_{\boldsymbol{\alpha}} \boldsymbol{m}_{\boldsymbol{\alpha}}(\boldsymbol{x}_i, t_i)]_{j'} - \boldsymbol{\mu}_{j'}\right) - \boldsymbol{\Sigma}_{j,j'}\right| \le t \ \forall j, j' && (61)\\
&\left|\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\right| \le t \ \forall j && (62)
\end{aligned}$$

On this event,

$$\begin{aligned}
\|\hat{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}^*) - \boldsymbol{\Sigma}\|_\infty &= \max_{j,j'}\left|\hat{\boldsymbol{\Sigma}}_{j,j'}(\boldsymbol{\alpha}^*) - \boldsymbol{\Sigma}_{j,j'}\right| \\
&\le \max_{j,j'}\left|\frac{1}{n}\sum_{i=1}^n \left([\nabla_{\boldsymbol{\alpha}} \boldsymbol{m}_{\boldsymbol{\alpha}}(\boldsymbol{x}_i, t_i)]_j - \boldsymbol{\mu}_j\right)\left([\nabla_{\boldsymbol{\alpha}} \boldsymbol{m}_{\boldsymbol{\alpha}}(\boldsymbol{x}_i, t_i)]_{j'} - \boldsymbol{\mu}_{j'}\right) - \boldsymbol{\Sigma}_{j,j'}\right| + \max_j\left|\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\right|^2 \\
&\le t + t^2 \le 2t
\end{aligned}$$

By above statement, we have by boundness of random variable $\nabla_{\boldsymbol{\alpha}} \boldsymbol{m}_{\boldsymbol{\alpha}}(\boldsymbol{x}, t)$ and Hoffeding's inequality, there exists c_1, c_2 depend on M only that

$$\begin{aligned}
\mathbb{P}\left(\left|\left([\nabla_{\boldsymbol{\alpha}} \boldsymbol{m}_{\boldsymbol{\alpha}}(\boldsymbol{x}_i, t_i)]_j - \boldsymbol{\mu}_j\right)\left([\nabla_{\boldsymbol{\alpha}} \boldsymbol{m}_{\boldsymbol{\alpha}}(\boldsymbol{x}_i, t_i)]_{j'} - \boldsymbol{\mu}_{j'}\right) - \boldsymbol{\Sigma}_{j,j'}\right| \ge t\right) &\le 2\exp\{-c_1 t^2 n\} && (63)\\
\mathbb{P}\left(\left|\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\right| \ge t\right) &\le 2\exp\{-c_2 t^2 n\} && (64)
\end{aligned}$$

Thus

$$\mathbb{P}\left(\|\hat{\mathbf{\Sigma}}(\boldsymbol{\alpha}^*) - \mathbf{\Sigma}\|_\infty \geq t\right) \leq 2k\exp\{-c_1 t^2 n\} + 2k^2\exp\{-c_2 t^2 n\} \leq 4k^2\exp\{-c_3 t^2 n\} \tag{65}$$

for some c_3 depend on M only. Finally we have

$$\mathbb{P}\left(\|\hat{\mathbf{\Sigma}}(\boldsymbol{\alpha}^*) - \mathbf{\Sigma}\|_\infty \geq t\right) \leq c\exp\{-c' t^2 n\} \tag{66}$$

by simplifying equation (65) with bound of t via choosing proper c_0 satisfying $c_0 > \sqrt{\frac{2}{c_3}}$. □

E.2 Inverse Hessian Approximation

Lemma E.4 (Consistency of Inverse Hessian estimator)**.** *Let $S_{\boldsymbol{\omega},j}$ be the support of $\boldsymbol{\omega}_j^*$, and $s_{\boldsymbol{\omega},j} = |S_{\boldsymbol{\omega},j}|$, under condition that*

$$\lambda_j \geq 2\|\nabla^2_{\boldsymbol{\alpha}}\hat{\mathcal{L}}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^*)\boldsymbol{\omega}_j^* - \boldsymbol{e}_j\|_\infty \tag{67}$$

we have

$$\|\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\|_1 \leq \frac{6}{\kappa_j} s_{\boldsymbol{\omega},j}\lambda_j \tag{68}$$

Proof. The proof is exactly the same as proof of Theorem D.1, the only difference is we replace $\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}(\boldsymbol{\alpha})$ with $\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^*)\boldsymbol{\omega}_j^* - \boldsymbol{e}_j$ and S with $S_{\boldsymbol{\omega},j}$. So we omit the proof. □

Lemma E.5. *There exists constants c_0, c, c' depend only on $M\|\boldsymbol{\omega}_j^*\| + 1$ such that for any $t \in [c_0\sqrt{\frac{\log k}{n}}, 1]$, we have*

$$\mathbb{P}\left(\|\nabla^2_{\boldsymbol{\alpha}}\hat{\mathcal{L}}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^*)\boldsymbol{\omega}_j^* - \boldsymbol{e}_j\|_\infty \geq t\right) \leq c\exp\{-c' t^2 n\} \tag{69}$$

Proof. The result follows from fact that the random variable $\nabla^2_{\boldsymbol{\alpha}}\hat{\mathcal{L}}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^*)\boldsymbol{\omega}_j^* - \boldsymbol{e}_j$ follows sub-Gaussian distribution by definition of $\boldsymbol{\omega}^*$. Therefore there exists c_0 such that

$$\mathbb{E}[\|\nabla^2_{\boldsymbol{\alpha}}\hat{\mathcal{L}}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^*)\boldsymbol{\omega}_j^* - \boldsymbol{e}_j\|_\infty] \leq c_0\sqrt{\frac{\log k}{n}} \tag{70}$$

The proof of equation (70) is the same as Lemma D.5. Finally we apply Hoffeding's inequality and obtain probabilistic inequality as desired. □

E.3 Gaussian Approximation Bound

This subsection presents the proof of Gaussian Approximation Bound(GAB). Lemma E.6 and E.7 are useful lemmas for the proof. Theorem E.8 talks about GAB.

Lemma E.6. *For $\boldsymbol{\omega} \in \mathbb{R}^k$, let*

$$A_n(\boldsymbol{\omega}) = \langle \boldsymbol{\omega}, \nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\rangle \tag{71}$$

and

$$\sigma_n^2 = \sigma_n^2(\boldsymbol{\omega}) = Var[\sqrt{n}A_n(\boldsymbol{\omega})] \tag{72}$$

Then

$$\sup_{z\in\mathbb{R}} |\mathbb{P}\{\sqrt{n}A_n(\boldsymbol{\omega})/\sigma_n \leq z\} - \Phi(z)| \leq \frac{2CM||\boldsymbol{\omega}||}{\sqrt{n}\sigma_n} \tag{73}$$

where $C = 3.3$ is a known constant.

Proof. We have

$$\frac{\sqrt{n}A_n(\boldsymbol{\omega})}{\sigma_n} = \frac{1}{\sqrt{n}}\left\{\sum_{i=1}^n \frac{\langle \boldsymbol{\omega}, \nabla_{\boldsymbol{\alpha}}\boldsymbol{m}_{\boldsymbol{\alpha}^*}(t_i, \boldsymbol{x}_i)\rangle}{\sigma_n}\right\} \tag{74}$$

and

$$\langle \boldsymbol{\omega}, \nabla_{\boldsymbol{\alpha}} \boldsymbol{m}_{\boldsymbol{\alpha}^*}(t_i, \boldsymbol{x}_i) \rangle \leq \frac{2M\|\boldsymbol{\omega}\|}{\sigma_n} \tag{75}$$

for all i. Finally the Berry-Esseen inequality (Theorem 3.4 of (Chen et al., 2010)) yields

$$\sup_{z \in \mathbb{R}} |\mathbb{P}\{\sqrt{n} A_n(\boldsymbol{\omega})/\sigma_n \leq z\} - \Phi(z)| \leq \frac{2CM||\boldsymbol{\omega}||}{\sqrt{n}\sigma_n} \tag{76}$$

where $C = 3.3$ is a known constant. □

Lemma E.7. *(Lemma D.3 of (Barber and Kolar, 2018)) If*

$$\sup_{z \in \mathbb{R}} |\mathbb{P}\{A \leq z\} - \Phi(z)| \leq \epsilon_A \text{ and } \mathbb{P}\{|B| \leq \delta_B, |C| \leq \delta_C\} \geq 1 - \epsilon_{BC} \tag{77}$$

for some $\epsilon_A, \epsilon_{BC}, \delta_B, \delta_C \in [0, 1)$, then

$$\sup_{z \in \mathbb{R}} |\mathbb{P}\{(A + B)/(1 + C) \leq z\} - \Phi(z)| \leq \delta_B + \frac{\delta_C}{1 - \delta_C} + \epsilon_A + \epsilon_{BC} \tag{78}$$

Theorem E.8. *(Restatement of Theorem 6.5) For $\delta_{\boldsymbol{\alpha}}, \delta_{\boldsymbol{\omega}}, \lambda_{\text{lasso}}, \lambda_j, \delta_\sigma \in [0, 1)$, define the event*

$$\boldsymbol{\mathcal{E}} = \left\{ \begin{array}{ll} 2\left\|\nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\right\|_\infty \leq \lambda_{\text{lasso}}, & 2\left\|\nabla^2_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\boldsymbol{\omega}_j^* - \boldsymbol{e}_j\right\|_\infty \leq \lambda_j, \\ \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 \leq \delta_{\boldsymbol{\alpha}}, & \left\|\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right\|_1 \leq \delta_{\boldsymbol{\omega}}, \\ ||\hat{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}^*) - \boldsymbol{\Sigma}||_\infty \leq \delta_\sigma/2 & \end{array} \right\} \tag{79}$$

Suppose $\mathbb{P}(\boldsymbol{\mathcal{E}}) \geq 1 - \epsilon$. We have

$$\sup_{z \in \mathbb{R}} \left| \mathbb{P}\left\{ \frac{\sqrt{n}(\tilde{\alpha}_j - \alpha_j^*)}{\hat{\sigma}_j} \leq z \right\} - \Phi(z) \right| \leq \Delta_1 + \Delta_2 + \Delta_3 + \epsilon \tag{80}$$

Proof. We have by definition of debiased lasso

$$\tilde{\alpha}_j = \hat{\alpha}_j - \tilde{\boldsymbol{\omega}}_j^\top \nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\hat{\boldsymbol{\alpha}}) \tag{81}$$

$$= \hat{\alpha}_j - \boldsymbol{\omega}_j^{*\top} \nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\hat{\boldsymbol{\alpha}}) + (\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*)^\top \nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\hat{\boldsymbol{\alpha}}) \tag{82}$$

$$= \hat{\alpha}_j - \boldsymbol{\omega}_j^{*\top} \left(\nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*) + \nabla^2_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \right) + (\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*) \nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\hat{\boldsymbol{\alpha}}) \tag{83}$$

$$\implies \tilde{\alpha}_j - \boldsymbol{\alpha}_j^* = \underbrace{-\boldsymbol{\omega}_j^{*\top} \nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)}_{A} \underbrace{- \left(\nabla^2_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\boldsymbol{\omega}_j^* - \boldsymbol{e}_j \right) (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) + (\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*)^\top \nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\hat{\boldsymbol{\alpha}})}_{B} \tag{84}$$

Therefore we have by lemma E.6 that

$$\sup_{z \in \mathbb{R}} |\mathbb{P}\{\sqrt{n} A/\sigma_n \leq z\} - \Phi(z)| \leq \frac{6.6M||\boldsymbol{\omega}_j^*||}{\sqrt{n}\sigma_j} = \Delta_1 \tag{85}$$

Furthermore we have

$$B = \left(\nabla^2_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\boldsymbol{\omega}_j^* - \boldsymbol{e}_j \right) (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) + (\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*)^\top \nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\hat{\boldsymbol{\alpha}}) \tag{86}$$

$$= \underbrace{\left(\nabla^2_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\boldsymbol{\omega}_j^* - \boldsymbol{e}_j \right) (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)}_{B_1} \tag{87}$$

$$+ \underbrace{(\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*)^\top \nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)}_{B_2} + \underbrace{(\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*)^\top \left(\nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\hat{\boldsymbol{\alpha}}) - \nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*) \right)}_{B_3} \tag{88}$$

and by Taylor expansion

$$\nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\hat{\boldsymbol{\alpha}}) - \nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*) = (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \frac{\tilde{\boldsymbol{F}}^\top G \tilde{\boldsymbol{F}}}{n} \tag{89}$$

Hence by Holder's inequality:

$$|B_1| \leq 2 \left\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\right\|_1 \left\|\nabla^2_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\boldsymbol{\omega}_j^* - \boldsymbol{e}_j\right\|_\infty \leq \lambda_j \delta_{\boldsymbol{\alpha}} \tag{90}$$

$$|B_2| \leq 2 \left\|\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right\|_1 \left\|\nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\right\|_\infty \leq \delta_{\boldsymbol{\omega}} \lambda_{\text{lasso}} \tag{91}$$

$$|B_3| \leq K \left\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\right\|_\infty \left\|\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right\|_1 \leq 4K \left\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\right\|_1 \left\|\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right\|_1 \leq K \delta_{\boldsymbol{\alpha}} \delta_{\boldsymbol{\omega}} \tag{92}$$

where $K = C_f^2 C_g$ is a constant and C_f, C_g are bounds of $||\boldsymbol{f}(\boldsymbol{x})||_\infty$ and $|g(t)|$ respectively, recall we assume bounded sufficient statistics. Therefore we have

$$\frac{\sqrt{n}|B|}{\sigma_j} \leq \frac{\sqrt{n}(\lambda_j \delta_{\boldsymbol{\alpha}} + \delta_{\boldsymbol{\omega}} \lambda_{\text{lasso}} + K \delta_{\boldsymbol{\alpha}} \delta_{\boldsymbol{\omega}})}{\sigma_j} = \Delta_2 \tag{93}$$

Moreover, by Lemma E.2,

$$C = \left|\frac{\hat{\sigma}_j}{\sigma_j} - 1\right| = \left|\frac{\hat{\sigma}_j - \sigma_j}{\sigma_j}\right| \leq \left|\frac{\hat{\sigma}_j^2 - \sigma_j^2}{\sigma_j^2}\right| \leq \frac{(2L\delta_{\boldsymbol{\alpha}} + \delta_\sigma)\left(\|\boldsymbol{\omega}_j^*\|^2 + \delta_{\boldsymbol{\omega}}^2\right) + |||\boldsymbol{\Sigma}|||_\infty \delta_{\boldsymbol{\omega}}^2}{\sigma_j^2} = \delta_C \tag{94}$$

where first inequality can be derived using the difference of squares formula. Finally we apply E.7, obtain

$$\sup_{z \in \mathbb{R}} \left|\mathbb{P}\{\sqrt{n}(\tilde{\alpha}_j - \alpha_j^*)/\hat{\sigma}_j \leq z\} - \Phi(z)\right| \leq \Delta_1 + \Delta_2 + \Delta_3 + \epsilon \tag{95}$$

where $\Delta_3 = \frac{\delta_C}{1-\delta_C}$. □

E.4 Proof of Theorem 6.7

Theorem E.9. *Denoting $s_{\boldsymbol{\omega},j}$ as the cardinality of support set of $\boldsymbol{\omega}_j^*$ and s as the cardinality of support set of $\boldsymbol{\alpha}^*$ respectively. Let $\tilde{\alpha}_j$ be debiased lasso estimator with tuning parameter*

$$\lambda_{\text{lasso}} \in \mathcal{O}\left(\sqrt{\frac{\log k}{n}}\right); \text{ and } \lambda_j \in \mathcal{O}\left(\sqrt{s_{\boldsymbol{\omega},j} \frac{\log k}{n}}\right) \tag{96}$$

we have there exists positive constants c, c' such that

$$\sup_{z \in \mathbb{R}} \left|\mathbb{P}\{\sqrt{n}(\tilde{\alpha}_j - \alpha_j^*)/\hat{\sigma}_j \leq z\} - \Phi(z)\right| \leq \mathcal{O}\left(\sqrt{n} s_{\boldsymbol{\omega},j}^{3/2} s \frac{\log k}{n}\right) + c \exp\{-c' \log k\} \tag{97}$$

Proof. Consider the event $\boldsymbol{\mathcal{E}}^L$

$$2||\nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)||_\infty \leq \lambda_{\text{lasso}} \tag{98}$$

$$2||\nabla^2_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\boldsymbol{\omega}_j^* - \boldsymbol{e}_j||_\infty \leq \lambda_j \tag{99}$$

$$||\hat{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}^*) - \boldsymbol{\Sigma}||_\infty \leq \sqrt{s}\lambda_{\text{lasso}} \tag{100}$$

We have $\boldsymbol{\mathcal{E}}^L \subseteq \boldsymbol{\mathcal{E}}$ under Assumption 6.3 since the following:
First we have by Theorem D.1 that from equation (98) with Assumption 6.3

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 \leq \frac{6}{\kappa}\lambda_{\text{lasso}} s \in \mathcal{O}\left(s\sqrt{\frac{\log k}{n}}\right) \tag{101}$$

In addition, we have from Lemma E.4 that from equation (99)

$$\|\tilde{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\|_1 \leq \frac{6}{\kappa}\lambda_j s_{\boldsymbol{\omega},j} \in \mathcal{O}\left(\sqrt{s_{\boldsymbol{\omega},j}^3 \frac{\log k}{n}}\right) \tag{102}$$

therefore $\boldsymbol{\mathcal{E}}^L \subseteq \boldsymbol{\mathcal{E}}$ and we have

$$\Delta_2 \in \mathcal{O}\left(\sqrt{n} s_{\boldsymbol{\omega},j}^{3/2} s \frac{\log k}{n}\right) \tag{103}$$

We ignore Δ_1 and Δ_3 since they are of smaller order.
Next we bound $\mathbb{P}\left(\{\boldsymbol{\mathcal{E}}^{\boldsymbol{L}}\}^c\right)$. Let

$$\boldsymbol{\mathcal{E}}_1 = \{2||\nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)||_\infty \leq \lambda_{\text{lasso}}\} \tag{104}$$

$$\boldsymbol{\mathcal{E}}_2 = \{2||\nabla^2_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)\boldsymbol{\omega}_j^* - \boldsymbol{e}_j||_\infty \leq \lambda_j\} \tag{105}$$

$$\boldsymbol{\mathcal{E}}_3 = \{||\hat{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}^*) - \boldsymbol{\Sigma}||_\infty \leq \sqrt{s}\lambda_{\text{lasso}}\} \tag{106}$$

It is obvious that

$$\mathbb{P}\left(\{\boldsymbol{\mathcal{E}}^{\boldsymbol{L}}\}^c\right) \leq \sum_{i=1}^{3} \mathbb{P}\left(\boldsymbol{\mathcal{E}}_i^c\right) \tag{107}$$

Under bounded condition, we have the following: By Lemma D.5 and Lemma E.5, there exist constants c_1, c_1', c_2, and c_2'.

$$\mathbb{P}\left(\boldsymbol{\mathcal{E}}_1^c\right) = \mathbb{P}\left(2||\nabla_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*)||_\infty > \lambda_k\right) \leq c_1 \exp\{-c_1' \log k\} \tag{108}$$

$$\mathbb{P}\left(\boldsymbol{\mathcal{E}}_2^c\right) = \mathbb{P}\left(2||\nabla^2_{\boldsymbol{\alpha}} \hat{\mathcal{L}}(\boldsymbol{\alpha}^*) - \boldsymbol{e}_k||_\infty > \lambda_{\text{lasso}}\right) \leq c_2 \exp\{-c_2' \log k\} \tag{109}$$

and by Lemma E.3 there exists c_3, c_3'

$$\mathbb{P}\left(\boldsymbol{\mathcal{E}}_3^c\right) = \mathbb{P}\left(||\hat{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}^*) - \boldsymbol{\Sigma}||_\infty > \sqrt{s}\lambda_{\text{lasso}}\right) \leq c_3 \exp\{-c_3' \log k\} \tag{110}$$

Therefore there exists c, c'

$$\mathbb{P}\left(\{\boldsymbol{\mathcal{E}}^{\boldsymbol{L}}\}^c\right) \leq c \exp\{-c' \log k\} \tag{111}$$

Finally we complete the proof with combining the bounds of (111) and (102),

$$\sup_{z \in \mathbb{R}} \left|\mathbb{P}\{\sqrt{n}(\tilde{\alpha}_j - \alpha_j^*)/\hat{\sigma}_j \leq z\} - \Phi(z)\right| \leq \mathcal{O}\left(\sqrt{n} s_{\boldsymbol{\omega},j}^{3/2} s \frac{\log k}{n}\right) + c \exp\{-c' \log k\} \tag{112}$$

□

F Score Model, a Gaussian Example

Let us group this example into three parts: fixed mean and time-dependent variance (μ and σ_t^2), time-dependent mean and fixed variance (μ_t and σ^2), and time-dependent mean and variance (μ_t and σ_t^2). Each one is detailed in distinct sections below.

Across all three cases we aim to verify the following equation holds

$$\partial_t \log q_t = \langle \partial_t \boldsymbol{\theta}(t), \boldsymbol{f}(x) - \mathbb{E}_{q_t}[\boldsymbol{f}(y)] \rangle, \tag{113}$$

which is our formulation given by Proposition 3.1. For the exponential family, the natural parameterisation of the Gaussian distribution is given by

$$\boldsymbol{\theta} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} \tag{114}$$

for a given μ and σ^2.

Fixed mean and time-dependent variance Let the fixed mean be written as μ and the time-dependent variance be written as σ_t^2. Firstly, according to this Gaussian distribution, the LHS of Equation (113) is given by

$$\begin{aligned}\partial_t \log q_t &= \partial_t \left(\frac{-(x-\mu)^2}{2\sigma_t^2}\right) - \partial_t \log(\sqrt{2\pi}\sigma_t) \\ &= -\partial_t \left(\frac{x^2}{2\sigma_t^2}\right) + \partial_t \left(\frac{x\mu}{\sigma_t}\right) - \partial_t \left(\frac{\mu^2}{2\sigma_t^2}\right) - \partial_t \left(\log \sigma_t\right) \\ &= -x^2 \partial_t \left(\frac{1}{2\sigma_t^2}\right) + x\mu \partial_t \left(\frac{1}{\sigma_t}\right) - \mu^2 \partial_t \left(\frac{1}{2\sigma_t^2}\right) - \frac{\partial_t \sigma_t}{\sigma_t}\end{aligned} \tag{115}$$

We aim to show that when $\boldsymbol{f}(x) = [x, x^2]$, the RHS of Equation (113) is equal to this. We first write

$$\boldsymbol{\theta}_t = \begin{bmatrix} \frac{\mu}{\sigma_t^2} \\ -\frac{1}{2\sigma_t^2} \end{bmatrix}, \qquad \partial_t \boldsymbol{\theta}(t) = \begin{bmatrix} \mu \partial_t \left(\frac{1}{\sigma_t^2}\right) \\ -\partial_t \left(\frac{1}{2\sigma_t^2}\right) \end{bmatrix}.$$

$$\begin{aligned}\langle \partial_t \theta_t, \boldsymbol{f}(x) - \mathbb{E}_{q_t}[\boldsymbol{f}(y)] \rangle &= \left\langle \begin{bmatrix} \partial_t \left(\frac{\mu}{\sigma_t^2}\right) \\ \partial_t \left(-\frac{1}{2\sigma_t^2}\right) \end{bmatrix}, \begin{bmatrix} x \\ x^2 \end{bmatrix} - \begin{bmatrix} \mathbb{E}_{q_t}[y] \\ \mathbb{E}_{q_t}[y^2] \end{bmatrix} \right\rangle \\ &= \left\langle \begin{bmatrix} \partial_t \left(\frac{\mu}{\sigma_t^2}\right) \\ \partial_t \left(-\frac{1}{2\sigma_t^2}\right) \end{bmatrix}, \begin{bmatrix} x \\ x^2 \end{bmatrix} - \begin{bmatrix} \mu \\ \sigma_t^2 + \mu^2 \end{bmatrix} \right\rangle \\ &= \left\langle \begin{bmatrix} \partial_t \left(\frac{\mu}{\sigma_t^2}\right) \\ \partial_t \left(-\frac{1}{2\sigma_t^2}\right) \end{bmatrix}, \begin{bmatrix} x - \mu \\ x^2 - \sigma_t^2 - \mu^2 \end{bmatrix} \right\rangle \\ &= \partial_t \left(\frac{\mu}{\sigma_t^2}\right)(x-\mu) - \partial_t \left(\frac{1}{2\sigma_t^2}\right)(x^2 - \sigma_t^2 - \mu^2) \\ &= x\mu \partial_t \left(\frac{1}{\sigma_t^2}\right) - \mu^2 \partial_t \left(\frac{1}{\sigma_t^2}\right) - x^2 \partial_t \left(\frac{1}{2\sigma_t^2}\right) + \sigma_t^2 \partial_t \left(\frac{1}{2\sigma_t^2}\right) + \mu^2 \partial_t \left(\frac{1}{2\sigma_t^2}\right) \\ &= x\mu \partial_t \left(\frac{1}{\sigma_t^2}\right) - \mu^2 \partial_t \left(\frac{1}{2\sigma_t^2}\right) - x^2 \partial_t \left(\frac{1}{2\sigma_t^2}\right) + \sigma_t^2 \partial_t \left(\frac{1}{2\sigma_t^2}\right).\end{aligned}$$

By the chain rule,

$$\sigma_t^2 \partial_t \left(\frac{1}{2\sigma_t^2}\right) = \frac{\sigma_t^2}{2} \partial_t \left(\sigma_t^{-2}\right) = -\frac{\sigma_t^2}{2} \frac{2}{\sigma_t^3} \partial_t \sigma_t = -\frac{\partial_t \sigma_t}{\sigma_t},$$

which, substituted into the equation above, leaves

$$x\mu \partial_t \left(\frac{1}{\sigma_t^2}\right) - \mu^2 \partial_t \left(\frac{1}{2\sigma_t^2}\right) - x^2 \partial_t \left(\frac{1}{2\sigma_t^2}\right) - \frac{\partial_t \sigma_t}{\sigma_t}$$

for which all terms match the terms in Equation (115) as desired.

Since we are doing very similar operations across all examples, the following two derivations will be lighter on details but should be straightforward to follow.

Time-dependent mean and fixed variance Firstly, write μ_t and σ^2 as the mean and variance of this Gaussian distribution, respectively. The time score function, i.e. the LHS of Equation (113) is given by

$$\begin{aligned}\partial_t \log q_t &= \partial_t \left(\frac{-(x-\mu_t)^2}{2\sigma^2}\right) - \partial_t \log(\sqrt{2\pi}\sigma) \\ &= \frac{2x\partial_t \mu_t - \partial_t(\mu_t^2)}{2\sigma^2}.\end{aligned} \tag{116}$$

The natural parameters are given by

$$\boldsymbol{\theta}_t = \begin{bmatrix} \frac{\mu_t}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}, \qquad \partial_t \boldsymbol{\theta}(t) = \begin{bmatrix} \partial_t \left(\frac{\mu_t}{\sigma^2}\right) \\ \partial_t \left(-\frac{1}{2\sigma^2}\right) \end{bmatrix} = \begin{bmatrix} \frac{\partial_t \mu_t}{\sigma^2} \\ 0 \end{bmatrix},$$

and so we consider the first dimension only. The RHS of Equation (113), when $\boldsymbol{f}(x) = x$ is given by

$$\begin{aligned}\langle \partial_t \theta_t, \boldsymbol{f}(x) - \mathbb{E}_{q_t}[\boldsymbol{f}(y)] \rangle &= \left\langle \frac{\partial_t \mu_t}{\sigma^2}, x - \mu_t \right\rangle \\ &= \frac{x\partial_t \mu_t - \mu_t \partial_t \mu_t}{\sigma^2} \\ &= \frac{x\partial_t \mu_t - \frac{\partial_t(\mu_t^2)}{2}}{\sigma^2} \\ &= \frac{2x\partial_t \mu_t - \partial_t(\mu_t^2)}{2\sigma^2}\end{aligned}$$

where the penultimate line is due to the chain rule. This matches Equation (116) as desired.

Time-dependent mean and variance Firstly, write μ_t and σ_t^2 as the mean and variance of this Gaussian distribution, respectively. The time score function, i.e. the LHS of Equation (113) is given by

$$\begin{aligned}\partial_t \log q_t &= \partial_t \left(\frac{-(x-\mu_t)^2}{2\sigma_t^2} \right) - \partial_t \log(\sqrt{2\pi}\sigma_t) \\ &= -x^2 \partial_t \left(\frac{1}{2\sigma_t^2} \right) + x \frac{\partial_t \mu_t}{\sigma_t^2} + x\mu_t \partial_t \left(\frac{1}{\sigma_t^2} \right) - \frac{\partial_t(\mu_t^2)}{2\sigma_t^2} - \mu_t^2 \partial_t \left(\frac{1}{2\sigma_t^2} \right) - \frac{\partial_t \sigma_t}{\sigma_t}. \end{aligned} \tag{117}$$

The natural parameters are given by

$$\boldsymbol{\theta}_t = \begin{bmatrix} \frac{\mu_t}{\sigma_t^2} \\ -\frac{1}{2\sigma_t^2} \end{bmatrix}, \qquad \partial_t \boldsymbol{\theta}(t) = \begin{bmatrix} \partial_t \left(\frac{\mu_t}{\sigma_t^2} \right) \\ \partial_t \left(-\frac{1}{2\sigma_t^2} \right) \end{bmatrix} = \begin{bmatrix} \frac{\partial_t \mu_t}{\sigma_t^2} + \mu_t \partial_t \left(\frac{1}{\sigma_t^2} \right) \\ -\partial_t \left(\frac{1}{2\sigma_t^2} \right) \end{bmatrix},$$

The RHS of Equation (113) when $\boldsymbol{f}(x) = [x, x^2]$ is given by

$$\begin{aligned}\langle \partial_t \theta_t, \boldsymbol{f}(x) - \mathbb{E}_{q_t}[\boldsymbol{f}(y)] \rangle &= \left\langle \begin{bmatrix} \frac{\partial_t \mu_t}{\sigma_t^2} + \mu_t \partial_t \left(\frac{1}{\sigma_t^2} \right) \\ -\partial_t \left(\frac{1}{2\sigma_t^2} \right) \end{bmatrix}, \begin{bmatrix} x - \mu_t \\ x^2 - \sigma_t^2 - \mu_t^2 \end{bmatrix} \right\rangle \\ &= \left(\frac{\partial_t \mu_t}{\sigma_t^2} + \mu_t \partial_t \left(\frac{1}{\sigma_t^2} \right) \right)(x - \mu_t) - \partial_t \left(\frac{1}{2\sigma_t^2} \right)(x^2 - \sigma_t^2 - \mu_t^2) \\ &= x\frac{\partial_t \mu_t}{\sigma_t^2} + x\mu_t \partial_t \left(\frac{1}{\sigma_t^2} \right) - \mu_t \frac{\partial_t \mu_t}{\sigma_t^2} - \mu_t^2 \partial_t \left(\frac{1}{\sigma_t^2} \right) - x^2 \partial_t \left(\frac{1}{2\sigma_t^2} \right) + \sigma_t^2 \partial_t \left(\frac{1}{2\sigma_t^2} \right) + \mu_t^2 \partial_t \left(\frac{1}{2\sigma_t^2} \right) \\ &= x\frac{\partial_t \mu_t}{\sigma_t^2} + x\mu_t \partial_t \left(\frac{1}{\sigma_t^2} \right) - \mu_t \frac{\partial_t \mu_t}{\sigma_t^2} + \mu_t^2 \partial_t \left(\frac{1}{2\sigma_t^2} \right) - x^2 \partial_t \left(\frac{1}{2\sigma_t^2} \right) + \sigma_t^2 \partial_t \left(\frac{1}{2\sigma_t^2} \right) \\ &= x\frac{\partial_t \mu_t}{\sigma_t^2} + x\mu_t \partial_t \left(\frac{1}{\sigma_t^2} \right) - \mu_t \frac{\partial_t \mu_t}{\sigma_t^2} + \mu_t^2 \partial_t \left(\frac{1}{2\sigma_t^2} \right) - x^2 \partial_t \left(\frac{1}{2\sigma_t^2} \right) - \frac{\partial_t \sigma_t}{\sigma_t} \\ &= x\frac{\partial_t \mu_t}{\sigma_t^2} + x\mu_t \partial_t \left(\frac{1}{\sigma_t^2} \right) - \frac{\partial_t(\mu_t^2)}{2\sigma_t^2} - \mu_t^2 \partial_t \left(\frac{1}{2\sigma_t^2} \right) - x^2 \partial_t \left(\frac{1}{2\sigma_t^2} \right) - \frac{\partial_t \sigma_t}{\sigma_t},\end{aligned}$$

where the last three equalities, in their respective order, are due to collecting like terms, the chain rule on the last term, and the chain rule again on the third term. This matches Equation (117), completing the proof.

G Simulation Study Implementation Details

G.1 Construction of Gaussian Graphical Models for Estimation

In Section 8.1, we consider two Gaussian Graphical Models.

Random Sine Gaussian Graphical Models refers to Gaussian Graphical Models whose edges that vary with time are sine functions of t and the edges that depend on t are randomly chosen with a Bernoulli distribution with probability 0.02. We set the diagonal element 2 and off-diagonal elements are

$$\Theta'_{i,j}(t) = \Theta'_{j,i}(t) = \begin{cases} 0.5 \cdot \sin(10t) \text{ w.p. } 0.02 \\ 0 \text{ w.p. } 0.98 \end{cases} \quad \text{for } i \in [20], j \neq i \tag{118}$$

Random Linear Gaussian Graphical Models refers to Gaussian Graphical Models whose edges that vary with time are sine functions of t and the edges that depend on t are randomly chosen with a Bernoulli distribution with probability 0.023. We set the diagonal element 1 and off-diagonal elements are

$$\Theta'_{i,j}(t) = \Theta'_{j,i}(t) = \begin{cases} 0.45t \text{ w.p. } 0.023 \\ 0 \text{ w.p. } 0.977 \end{cases} \quad \text{for } i \in [40], j \neq i \tag{119}$$

G.1.1 Construction of Θ_0

To ensure the positive definiteness and symmetry of the matrix Θ_0, our procedure for constructing Θ_0 is as follows: First, we sample a matrix $A \in \mathbb{R}^{k\times k}$, where each element $A_{i,j} \sim \mathcal{N}(0,1)$. We then compute $A^\top A/d/2$. Finally, we replace the diagonal entries of $A^\top A$ with 0 and obtain Θ_0.

G.2 Construction of Gaussian Graphical Models for Inference

In Section 8.2, we considered two different types of Gaussian Graphical Models.

Deterministic Gaussian Graphical Models refers to a Gaussian Graphical Model whose edges vary linearly with time and those edges are manually chosen. In our experimental setting, we set the diagonal elements above and below the main diagonal of $\Theta'(t)$ to vary with time linearly, as well as the edge between the first node and third and forth node, specifically

$$\Theta'_{i,i+1}(t) = \Theta'_{i+1,i}(t) = t \text{ for } 1 \leq i \leq 19 \text{ and } \Theta'_{1,i}(t) = \Theta'_{i,1}(t) = t \text{ for } i = 3,4,5 \tag{120}$$

and remaining elements are set at 0.

Random Gaussian Graphical Models refers to a Gaussian Graphical Model whose edges vary with time linearly are random. In our experiment setting, we set the off-diagonal elements except edge of interest to follow a Bernoulli distribution with probability 0.2, i.e.

$$\Theta'_{i,j}(t) = \Theta'_{j,i}(t) = \begin{cases} t \text{ w.p. } 0.2 \\ 0 \text{ w.p. } 0.8 \end{cases} \quad \text{for } i \in [20], j \neq i \tag{121}$$

We then refill the diagonal entries to 0.

G.2.1 Construction of Θ_0

To ensure the positive definiteness and symmetry of the matrix Θ_0, our procedure for constructing Θ_0 is as follows: First, we sample a matrix $A \in \mathbb{R}^{k\times k}$, where each element $A_{i,j} \sim U(0,1)$ are sampled uniformly. We then compute $\delta A^\top A$, where we set $\delta = 0.01$. Finally, we replace the diagonal entries of $\delta A^\top A$ with 12 and obtain Θ_0, where the choice of 12 ensures the positive definiteness of $\Theta(t)^{-1}$ in the power test experiments.

G.3 Weighting Function

In Section 4.1, we specify the weight function $g(t)$ with the condition that it equals zero at the boundaries of the time domain. For truncated score matching, Liu et al. (2022) propose a distance function as g, from which we take inspiration. Let t_{start} and t_{end} denote the start and end of the time domain respectively. We propose

$$g(t) \coloneqq -(t - t_{\text{start}})(t - t_{\text{end}}), \tag{122}$$

and consequently $\partial_t g_t = -(2t - t_{\text{start}} - t_{\text{end}})$. In experimental results, we have observed that the choice of g does not have significant impact on the performance.

G.4 Hyperparameter Choice

In the coverage experiments, we set the regularization parameter for Lasso to be $\lambda_{\text{lasso}} = \sqrt{2\frac{\log k}{n}}$, and we use the same value for the regularization parameter in the inverse Hessian estimation.

G.5 Loggle for Testing Changing Edge

Loggle(Yang and Peng, 2020), built as an R package, is the main method we compared with in both estimation and inference task. We generate data and call the R package of Loggle from python.

The main challenge we face when implementing the Loggle in comparison is how to turn $\Theta(t)$ obtained by Loggle into $\partial_t \Theta(t)$ which is related to the change of edge. Here we use a heuristic approach, permutation tests to find appropriate quantiles for deciding whether the edge should be considered changing. We shuffle the data matrix 100 times so that the dependency between $\boldsymbol{x}$ and t are broken. Then we apply Loggle to the shuffled data and use least square linear regression to obtain the slope. The 2.5% and 97.5% quantiles of the set of slopes are set as thresholds; any slope from original data that falls outside the 95% coverage indicates a changing edge. When calculating the coverage and power, we compute new quantiles for each trial individually.