# ADVANCED CLASSIFICATION METHODS FOR LARGE SPATIAL-TEMPORAL DATA: APPLICATIONS TO NEUROIMAGING

By

Rejaul Karim

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics - Doctor of Philosophy

2019

ABSTRACT

ADVANCED CLASSIFICATION METHODS FOR LARGE SPATIAL-TEMPORAL DATA: APPLICATIONS TO NEUROIMAGING

By

Rejaul Karim

Spatial data are characterized by dependency between the data indexed by a fixed point in space and its "neighbors". Exploiting such dependencies leads to improvement in estimation and inference. Due to large abundance of such data in nature,previous methodologies are being extended to incorporate such proximal information.For example in a latent model for generating data is spatially dependent,one would like to investigate how such dependencies affect the variable selection performances. This work is centerd around a penalized estimating equation approach to model of an expanding dimension($p_n$) of predictor variables where responses are generated from Poisson model driven by latent Gaussian model (Log Gaussian Cox process). In the past this approach has been extensively studied in longitudinal data analysis. Gaussian random fields that exhibit Conditional autoregressive structure (CAR) we provide some theoretical results of the estimator obtained from the penalized estimating equation. The oracle properties of the estimator as described by Fan & Li (2001) are provided.

Pattern detection in imaging data has lead to a rise of classification methods that are effective in separating objects and structures in an image. This provides a major impetus in the context of medical imaging. Magnetic resonance images (MRI) collected in four dimensions (3D space and time), maybe used to predict different disease phases of a particular patient. Linear discriminant analysis(LDA) is a classical tool used for dimension reduction as well as classification. However, in the context of high-dimensional data where feature

volume is significantly larger than sample size, the within-class covariance of the LDA tool is singular, yielding the classification rule unsuitable. Sparse discriminant methods have therefore been proposed to implement LDA in a high dimensional setup. These methods do not incorporate dependencies in the feature covariance structure when data acquired is spatially and temporally correlated. This article proposes a regularized high dimensional LDA resolution for spatio-temporal imaging data. Theoretically we ensure that the method proposed can achieve consistent parameter estimation, feature selection, at an asymptotically optimal misclassification rate. Extensive simulation study shows a significant improvement in classification performance under spatial-temporal dependence. This method is applied to longitudinal structural MRI data obtained from the ADNI initiative.

LDA classification rule are restrictive since this paradigm is based on strong assumption that binary class data generating process is Normally distributed with same covariance function. In contrast, support vector machine is considered much more robust classifier due to its distribution free approach. The tensor counterpart of SVM also known as support tensor machine is widely popular in analysis of MRI image which is a tensor in its original format. This tensor structure preserves the neighboring spatial information which is lost after vectorization. In this work, we apply memory efficient random projection to tensor as a dimension reduction method which preserves distance with high probability. Near optimal classification consistency is shown along with few simulation study.

To the memory of my family members who have motivated and supported me.

## ACKNOWLEDGMENTS

I would like to sincerely thank my chair advisor Professor Tapabrata(Taps) Maiti for everything. Few words does not do justice to his contribution in my career. His patience and support resuscitated my academic career from time to time. I have utmost gratitude to Professor Chae Young Lim who spent lot of effort in proof checking and guided me through numerous expert suggestions. Her smart ideas helped me steer through research bottlenecks.

I will be indebted to my committee member Professor Yimin Xiao who introduced to spatial statistics. His mathematical intuition still amazes me. I am grateful to generosity of my Professor Arun Ross for his unconditional acceptance to be part of my committee. He inspires me through his interesting application approaches which revolutionized biometrics.

A very special mention to my senior and collaborator Abdhi Sarkar who helped me with superb coding skills especially regarding MRI data preprocessing and analysis. I would be unable to complete this dissertation without her. I am also thankful to my colleague Peide Li(Peter) who responded immediately to distress calls. The third work of my thesis is largely based on extension of his work. I am grateful to Yingjie Li who has permitted to use her material in my second work. The third work of my thesis is an extension for her thesis.
Lastly I would like to thank my parents and my sister for the moral support.

# TABLE OF CONTENTS

# LIST OF TABLES

# KEY TO SYMBOLS

1. $\leq$ denotes the positive semidefinite ordering. i.e. we write $A \leq B$ if the matrix $B - A$ is positive definite.
2. $|\cdot|$ denotes the cardinality of a set.
3. $\bigtriangledown_\beta \mathcal{K}$ is the gradient of a vector $\mathcal{K}$ w.r.t $\beta$.
4. $\circ$ signifies the Schur or Hadamard product of two matrices.
5. $\otimes$ signifies the Kronecker product of two matrices.
6. $\ddot{\mu}$ denotes the double derivative of the function $\mu$, similarly $\dddot{\mu}$ denotes the third derivative and so on.
7. $\coloneqq$ signifies assignment or is referred to as "denoted by".
8. Tr $A$ denotes the trace of a matrix A.
9. LHS stands for Left Hand Side of an equation.
10. $\sim$ denotes neighbors. i.e. $u \sim v$ implies $u$ and $v$ are adjacent voxels.
11. $\langle \cdot, \cdot \rangle$ signifies inner product of two vectors.
12. $sign(a)$ signifies $-1$ or $+1$ depending on whether $a < 0$ or $a > 0$ respectively.
13. $\mathbb{I}(\cdot)$ denotes the indicator operator.
14. TP (True Positives) represents the average number of correctly detected nonzero coefficients.
15. FP (False Positives) represents the average number of incorrectly detected nonzero coefficients.
16. CP gives the average empirical Coverage Probability of the 95% confidence intervals.
17. OSE stands for One step Estimation

# Chapter 1

# Analysis of Spatial Count Data: Penalized Estimating Equation Approach

## 1.1 Introduction

Geographical factors play a significant role in epidemiology. Poisson regression is popularly used for the analysis of disease rates, plant growth etc which assumes that the rates in nearby regions are independent and the variance of response is equal to the mean.Yasui & Lele (1997) Hierarchical models have been proposed to utilize spatial locations and neighbors as analysis of disease rates. Using hierarchal model, marginal likelihood becomes intractable, so this issue can be solved using estimating equation.

## 1.2 Literature review

Mardia and Marshall(stationary setup CAR low dimensional estimation) have produced consistent estimators in Gaussian CAR model using likelihood method.However since likelihood

is not tractable for Poisson log Normal Distribution. Mean field method cannot be used here it assumes the distribution of latent variable to independent. Following Yasui & Lele (1997) method, we need to find exact form of function $Y^*$ which is unbiased estimator for $\log Y$ to prove theoretical result. In order to achieve one way is to solve the first order Fredholm equation. However this equation is not solvable since $\log Y$ do not have finite expectation when $Y \sim \text{Poisson}(\lambda)$ Liang and Zeger (1986) used the idea of estimating equation for longitudinal data.Wang L(2011) extended this for diverging number of covariates. Wang L, Zhou J, Qu A.(2011) used penalized estimating equation for high dimensional inference. ClS type estimates and MM estimates and estimating equation approach is used. The central limit theorem for this statistics can be achieved by result using Peligard of rho mixing . Correlation upper bound is obtained for banded precision matrix. For general CAR model, problem remains open. Some progress can be made using partitioned matrix inverse $\Sigma_{ii}^{-1} = (\Sigma_{ii})^{-1} - \Sigma_{-i,i}(\Sigma_{-(i,i)})^{-1}\Sigma_{i,-i}$ LIN and CLAYTON (2005) extended the use estimating equation for spatial binary data referencing to the work of Zeger(1988) for time series of count data. These works shows that estimating equation approach can be used for spatially dependent data under appropriate mixing conditions. These paper estimated the regression parameter from quasi likelihood score function. Asymptotic covariance of such estimator depends on unknown nuisance parameter like scale and correlation parameters $(\sigma^2, \gamma)$ .Lele(1991),used Jacknifing tools for reduced bias estimators of nuisance parameters. HEAGERTY and LUMLEY (2000) proposed non parametric estimation of Covariance matrix using sub sampling windows for time series in general lattice data. Prentice(1988) provided a consistent estimate of scale and correlation parameter using second quasi likelihood score function. Here the precision matrix in quasi score requires knowledge of third and fourth moments which replaced by "working" precision matrix. This paper generalizes the

idea of joint optimal estimation Godambe and Thompson (1989) $(\beta_n, \sigma^2, \gamma)$ in independent when skewness and kurtosis of distribution is known from before.

## 1.3 Model

For county $i \in \{1, 2, .., n\}$ $Y_i$= Observed disease cases of county $i$. $E_i$= Expected disease cases of county $i$.$E_i$ s are known. $\Psi_i$= Logarithm of ratio of disease rates to some reference rates county $i$ .

$$Y_i|\Psi_i \overset{ind}{\sim} Poisson(e^{\Psi_i} E_i)$$

$$\boldsymbol{\psi}_{n\times 1} \sim \mathbb{N}\Big(\mathbb{X}_{n\times p_n}\boldsymbol{\beta}_{p_n\times 1}, \boldsymbol{\sigma}^2 \boldsymbol{V}_n(\boldsymbol{\gamma})\Big)$$

$$\boldsymbol{V}_n(\boldsymbol{\gamma}) = (\mathbb{I} - \boldsymbol{\gamma} M_n W_n)^{-1} M_n$$

$$\boldsymbol{\gamma} \in (-1, 1)\,;\, \boldsymbol{\sigma}^2 \in (0, \infty)$$

Define $\boldsymbol{W}_{n\times n}$ as adjacency matrix

$$W_{ij} = \begin{cases} 1 & \text{if i and j are neighbours} \\ 0 & \text{otherwise} \end{cases}$$

$\sum_{j\in\mathcal{N}(i)} W_{ij} = W_{i+}$. Let $\boldsymbol{M}_{n\times n} = \text{Diag } (\frac{1}{W_{i+}})_{ii}$

Here matrix $\boldsymbol{W}_n$ can be interpreted as the adjacency matrix of graph with vertex set as indices of random variable namely={1,2,...n}. $\boldsymbol{W}_n^{r\,i,j}$ represents the number of paths of length $r$ from vertex $i$ to $j$. $\boldsymbol{W}_n$ is irreducible if it is adjacency matrix of connected graph.Mathematically, $\boldsymbol{W}_n^{r\,i,j} > 0$ for some $r$ and all $i, j \in \{1, 2, .., n\}$. Suppose $\boldsymbol{W}_n$ is reducible then $\boldsymbol{W}_n$ can be represented into block diagonal matrix of irreducible matrix. The block diagonal structure of $\boldsymbol{W}_n$ represents the isolated connected components of corresponding graph. Suppose

$$\boldsymbol{W} = \begin{pmatrix} {}^1W & & & 0 \\ & {}^2W & & \\ & & \ddots & \\ 0 & & & {}^{c-1}W \\ & & & & {}^cW \end{pmatrix}$$

here matrix ${}^jW$ is irreducible adjacency matrix of graph with vertex set indexed by $j$

So the corresponding $\boldsymbol{V}_{1,n} = (\mathbb{I} - \boldsymbol{\gamma} M_n W_n)^{-1} M_n$ is also block diagonal matrix. Without loss of generality we prove all results assuming $\boldsymbol{W}_n$ is irreducible since these results can be extended to block diagonal $\boldsymbol{W}_n$ similarly.

## 1.4 Estimation

The objective is to estimate $\beta$,$\sigma^2$ and $\gamma$ using observations $\{Y\}_i$. Most straightforward method would me Maximum Likelihood Method (MLE) method but MLE is untractable so Yasui & Lele (1997) have used transformation and then MLE.

### 1.4.1 Transformation approach by Yasui & Lele (1997)

Define

$$Y^* = f_1(Y) \text{ such that } \quad \mathbb{E}_{Y^*|\Psi} = \Psi + b_*(\psi)$$

$$Y^{**} = f_2(Y) \quad \text{such that} \quad \mathbb{E}_{Y^{**}|\Psi} = \Psi^2 + b_{**}(\psi)$$

With $b_*(y)$ and $b_{**}(y)$ satisfying some regularizing condition. $\mathbf{q}(\boldsymbol{\lambda}_n)$ is SCAD penalty function with parameter $\boldsymbol{\lambda}_n$ , the following sets of MLE equation are solved for estimation

$$\mathbb{U}_n(\boldsymbol{\beta}_n) = \underbrace{\frac{1}{n}\mathbb{X}^\intercal\left[\boldsymbol{\sigma}^2\mathbf{V}_{\mathbf{3},\mathbf{n}}(\boldsymbol{\gamma})\right]^{-1}(\boldsymbol{Y}^* - \mathbb{X}\boldsymbol{\beta})}_{S_n} - \mathbf{q}(\boldsymbol{\lambda}_n)(\boldsymbol{\beta}_n)$$

$$\mathbb{F}_{1,n}(\boldsymbol{\gamma}) = (\boldsymbol{Y}^* - \mathbb{X}\boldsymbol{\beta})^\intercal \boldsymbol{M}_n\boldsymbol{W}_n(\mathbb{I} - \boldsymbol{\gamma}\boldsymbol{M}_n\boldsymbol{W}_n)(\boldsymbol{Y}^* - \mathbb{X}\boldsymbol{\beta}) = 0$$

$$\mathbb{F}_{2,n}(\boldsymbol{\sigma}^2) = (\boldsymbol{Y}^* - \mathbb{X}\boldsymbol{\beta})^\intercal(\boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\boldsymbol{W}_n)(\boldsymbol{Y}^* - \mathbb{X}\boldsymbol{\beta}) - n\boldsymbol{\sigma}^2 = 0$$

#### 1.4.1.1 High dimensional curse

These equations are derived assuming asymptotic unbiasedness of $\boldsymbol{Y}^*$ for large value of $\boldsymbol{\Psi}$ and plugin method of moment estimates i.e $\boldsymbol{Y}^*$ in place of $\boldsymbol{\Psi}$ in the estimating equations.BuT when $p_n$ is more than n, the above method breaks down due to accumulation of bias of order $p_n/n$. Shown in appendix.

## 1.5 In search for transformation function

Here we considered two approaches to obtain a function which can serve as unbiased estimate of $\mathbb{X}\beta$. Using these two method we could only approximate the target function. But high dimension such approximation fail miserably.

### 1.5.1 Integral equation approach

We want to solve for function $Y* = f(Y)$ such that $\mathbb{E}_{Y^*|\Psi} = \Psi$ or equivalently

$$\sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} f(i) = \log \lambda$$

for all $\lambda \in (0, \infty)$ This kind of equation is classified as Fredholm Integral equation of first kind with form

$$Kf = g$$

where $K$ is integral or expectation operator of random variable Poisson $\lambda$ with kernel $K(\lambda, i) = \frac{\lambda^i}{i!}$ . $g$ is known (data function) here $g(\lambda) = \log \lambda$ and $f$ (solution function) is to be solved for.
A natural heuristic is to think $K$ as matrix with rows, dependent on $\lambda$ and columns dependent upon $i \in \{0, 1, .., \infty\}$ where $f$ is an unknown vector and $g$ is constant depending on $\lambda$

Assuming all regularity conditions for Poisson kernel. Suppose it has SVD(singular value decomposition) the equation can be solved. But the data function $g(\lambda) = log(\lambda) \notin \mathbf{L}_1(\mathcal{L} = \text{Lebesgue}, \mathbb{R}, \mathcal{B})$. Hence $log(\lambda)$ is not complete with respect to orthonormal basis. Therefore it does not have a series expansion with respect to any orthonormal basis.

#### 1.5.1.1 Charlier Polynomial

Charlier polynomials $\{C_m\}_{m=0}^{\infty}$ are a family of orthogonal polynomials with respect to Poisson weights.

$$\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} C_m(\lambda, i)\, C_n(\lambda, i) = \lambda^{-n} e^{\lambda} n! \delta_{mn} \quad \lambda > 0$$

Suppose

$$\log(\lambda) = \sum_{n=0}^{\infty} a_n \lambda^{-n} n! \quad \text{for some constants } a_n$$

Then function $f(y)$ could be analytically approximated. However $\log(\lambda)$ does not have Taylor series expansion on $(0, +\infty)$. The solution $f$ cannot be found using Charlier Polynomial.

However many numerical approximations are possible using techniques like quadratures etc.

### 1.5.2 Properties of marginal model

Finally we came across the work of which gives basic ideas about marginal distribution of this hierarchal model also known as log poisson Model. We derive first few moments needed in order to solve for this estimating equation.

The induced model is $\mathbf{P\Lambda^{d=1}}(\mathbb{X}\boldsymbol{\beta}, \boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma}))$ multivariate Poisson-log normal distribution with moments provided in the appendix

## 1.6 Penalized quasi likelihood on Induced model

Basawa have discussed the likelihood estimation of int erst parameter under mixture model where the nuisance parameter has prior. He showed that under exponential family set up the conditional likelihood estimation is asymptotically efficient as the mixture distribution set up. However our problem does not fit this paradigm since the prior is on the random variable has prior on it.

Extending the idea of conditional least squares,it can be shown that there are two cases for estimation.In case 1 we can estimate all parameters jointly $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}^2$ by either by differentiating same pseudo likelihood with respect to different parameters. $h_i = \left(\boldsymbol{Y}_i - \mathbb{E}(\boldsymbol{Y}_i|\mathcal{F}_{-i})\right)^2$ then multiply appropriate weights for optimality.Here $\mathcal{F}_{-i} = (S)(\boldsymbol{Y}_1, ..., \boldsymbol{Y}_{i-1}, \boldsymbol{Y}_{i+1}, ...\boldsymbol{Y}_n)$ is Sigma field of all samples from 1 to $n$ but $\boldsymbol{Y}_i$ In case 2 we can use two different pseudo likelihoods i.e $h_i = \left(\boldsymbol{Y}_i - \mathbb{E}(\boldsymbol{Y}_i|\mathcal{F}_{-i})\right)^2$ and $h_i' = \left(\left(\boldsymbol{Y}_i - \mathbb{E}(\boldsymbol{Y}_i|\mathcal{F}_{-i})\right)^2 - Var(Y_i|\mathcal{F}_{-i})\right)^2$ then multiply appropriate weights for optimality. In our setup the expression of conditional expectation $\mathbb{E}(\boldsymbol{Y}_i|\mathcal{F}_{-i})$ is intractable

We work for simplified case: estimate all parameters through $h_i = (\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_n, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})^2$ the multiply appropriate weights for optimality.

## 1.7 GEE

Since we have few moments of marginal distribution we can formulate the estimating equation

Let $\mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^\intercal \\ \boldsymbol{X}_2^\intercal \\ \cdots \\ \boldsymbol{X}_n^\intercal \end{pmatrix}$ be the coefficient matrix for $\boldsymbol{\beta}$ Hence $\mathbf{X}_i^T$ is the $i^th$ row of $\mathbf{X}$

$$\boldsymbol{\mu}_i(\boldsymbol{\beta}_n, \boldsymbol{\sigma}_2, \boldsymbol{\gamma}) = \mathbb{E}(\boldsymbol{Y}_i) = exp\left(-\boldsymbol{X}_i^\intercal \boldsymbol{\beta}_n + \boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2\right) = \theta_{n,i} R_{ii}$$

$$V_n(\boldsymbol{\sigma}^2, \boldsymbol{\gamma}, \boldsymbol{\beta}_n) = Diag(\boldsymbol{\mu}_i) \; + \; Diag(\boldsymbol{\mu}_i)[\boldsymbol{R}^{\odot 2}(\boldsymbol{\sigma}^2, \boldsymbol{\gamma}) - \boldsymbol{J}_n]Diag(\boldsymbol{\mu}_i)$$

$$exp\left(-\boldsymbol{X}_i^\intercal \boldsymbol{\beta}_n\right) = \theta_{n,i}$$

$$exp(\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) = R_{ii}$$

$$\left(\boldsymbol{R}^{\odot 2}(\boldsymbol{\sigma}^2, \boldsymbol{\gamma}) - \boldsymbol{J}_n\right)_{i,j} = exp(\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n\ i,j}) - 1$$

$\boldsymbol{R}^{\odot 2}$ denotes Hadamard product of matrix $\boldsymbol{R} \odot \boldsymbol{R}$ and $\boldsymbol{J}_n = \mathbf{1}_n \mathbf{1}_n^\intercal$

The equation are derived from stationarity of KKT conditions. The following score functions arise when we treat $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}^2$ as mean parameters under the constraints

- sparse solution inducing penalty on $\boldsymbol{\beta}$

- $\boldsymbol{\sigma}^2 \geq 0$

- $|\boldsymbol{\gamma}| \leq 1$

1

$$\mathbb{U}_n\big(\boldsymbol{\beta}_n \boldsymbol{\sigma}^2, \boldsymbol{\gamma}\big) = \underbrace{\frac{1}{n}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}^\intercal \left[\mathbf{V_n}(\boldsymbol{\sigma^2}, \boldsymbol{\gamma}, \boldsymbol{\beta_n})\right]^{-1}\Big(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}_n, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})\Big)}_{S_n} + \mathbf{q}(\boldsymbol{\lambda}_n)(|\boldsymbol{\beta}_n|) \odot sgn(\boldsymbol{\beta}_n)$$

$$= \underbrace{-\frac{1}{n}\Big(\mathbf{X}_1\boldsymbol{\mu}_1, \quad \ldots, \mathbf{X}_n\boldsymbol{\mu}_n\Big)\left[\mathbf{V_n}(\boldsymbol{\sigma^2}, \boldsymbol{\gamma}, \boldsymbol{\beta_n})\right]^{-1}\Big(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}_n, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})\Big)}_{S_n} + \mathbf{q}(\boldsymbol{\lambda}_n)(|\boldsymbol{\beta}_n|) \odot sgn(\boldsymbol{\beta}_n)$$

$$= \underbrace{-\frac{1}{n}\mathbb{X}^T \mathbf{Diag}\big(\boldsymbol{\mu}_i\big)\left[\mathbf{V_n}(\boldsymbol{\sigma^2}, \boldsymbol{\gamma}, \boldsymbol{\beta_n})\right]^{-1}\Big(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}_n, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})\Big)}_{S_n} + \mathbf{q}(\boldsymbol{\lambda}_n)(|\boldsymbol{\beta}_n|) \odot sgn(\boldsymbol{\beta}_n)$$

$$= 0$$

under $\boldsymbol{\lambda}_n \geq 0$

$\mathbf{q}(\boldsymbol{\lambda}_n)$ is derivative of SCAD penalty function wrt $\boldsymbol{\beta}$ with penalty parameter $\boldsymbol{\lambda}_n$

2

$$\mathbb{F}'_{3,n}(\boldsymbol{\gamma}, \boldsymbol{\beta}_n, \boldsymbol{\sigma}^2) = \underbrace{\frac{1}{n}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\gamma}}^{\intercal}\left[\mathbf{V_n}(\boldsymbol{\sigma^2}, \boldsymbol{\gamma}, \boldsymbol{\beta_n})\right]^{-1}\Big(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}_n, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})\Big)}_{\mathbb{F}_{3,n}} + v\, sgn(\boldsymbol{\gamma})$$

$$= \frac{\boldsymbol{\sigma}^2}{2n}\ \boldsymbol{\mu}^{\intercal}\ \mathbf{Diag}(\boldsymbol{V}^{-1}_{\mathbf{1,n}}\boldsymbol{W}_{\mathbf{n}}\boldsymbol{V}^{-1}_{\mathbf{1,n}}(\boldsymbol{\gamma}))\ \left[\mathbf{V_n}(\boldsymbol{\sigma^2}, \boldsymbol{\gamma}, \boldsymbol{\beta_n})\right]^{-1}\Big(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}_n, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})\Big) + v\, sgn(\boldsymbol{\gamma})$$

$$= 0$$

under $v \geq 0$

3

$$\mathbb{F}'_{2,n}(\boldsymbol{\sigma}^2, \boldsymbol{\gamma}, \boldsymbol{\beta}_n) = \underbrace{\frac{1}{n}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\sigma^2}}^{\intercal}\left[\mathbf{V_n}(\boldsymbol{\sigma^2}, \boldsymbol{\gamma}, \boldsymbol{\beta_n})\right]^{-1}\Big(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}_n, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})\Big)}_{\mathbb{F}_{2,n}} - u$$

$$= \frac{1}{2n}\boldsymbol{\mu}^{\intercal}\ \mathbf{Diag}(\boldsymbol{V}_{\mathbf{1,n}}(\boldsymbol{\gamma}))\left[\mathbf{V_n}(\boldsymbol{\sigma^2}, \boldsymbol{\gamma}, \boldsymbol{\beta_n})\right]^{-1}\Big(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}_n, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})\Big) - u$$

$$= 0$$

under $u \geq 0$

#### 1.7.0.1 Short range dependence

We need short range dependence property which gurantees variable selection consistency $\boldsymbol{\sigma}^2, \boldsymbol{\gamma}, \boldsymbol{\beta}_n)$ i.e variance covariance matrix of $\boldsymbol{Y}$ defined above has finite row (or colomn) sum bounded from above and below. Here $\boldsymbol{V}_{1,n\ i,j} = f(\boldsymbol{V}_{n\ i,j})$ where $f : \mathbb{R} \to \mathbb{R}$ is hadamard matrix function

**Lemma 1.7.1.** $\|V_n\|_\infty \geq \sup_i(\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^2\boldsymbol{\sigma}^2\frac{1}{w_{i+}(1+|\boldsymbol{\gamma}|)}\ )$

**Lemma 1.7.2.** $\|V_n\|_\infty \leq \sup_i \frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^2\left(exp(\boldsymbol{\sigma}^2\boldsymbol{V}_{1,n\ i,i}) - 1\right)}{\boldsymbol{\sigma}^2\boldsymbol{V}_{1,n\ i,i}}\ \frac{1}{w_{i+}(1-|\boldsymbol{\gamma}|)}$

Since $\|V_n\|_\infty < \infty$ for all $n$ it implies $\sum_{j=1}^{\infty} COV(Y_i, Y_j) < \infty$ which represents short range dependence.

## 1.8 Penalty

To proof consistency penalized score function needs Taylor expansion. So unbiased penalty function is desired by Heyde constrained equation

### 1.8.1 Smooth Clipped Absolute Deviation Penalty

Non concave penalty with oracle properties by Fan and Li (2001)

$$q_{\lambda}^{'}(|\theta|) = \lambda[\mathbb{I}(|\theta| < \lambda) \ + \ \frac{(a\lambda - |\theta|)_{+}}{(a-1)\lambda}\mathbb{I}(|\theta| \geq \lambda)]$$

$a > 2$ for computation $a = 3.7$

Alternative penalty can be MCP.

$$q_{\lambda}^{'}(|\theta|) = \lambda(1 - \frac{|\theta|}{a\lambda})_{+}$$

$a > 0$ value for a are found using cross validation

Here $S_n(\hat{\boldsymbol{\beta}}_n)$ is non penalized score version . Then we add penalty term and prove asymptotic up crossing ie $U_n = o_p(a_n)$ where $a_n \to 0$ is sufficient since $U_n$ may be discontinuous due to penalty.

## 1.9 Assumptions

1. $E_i = 1$ for all $i$
2. True parameter $\boldsymbol{\beta}^{\intercal}_{n_{(1\times s_n)}} = (\boldsymbol{\beta}^{\intercal}_{n_{(1\times p_n)}}, \mathbf{0}_{(1\times s_n - pn)})$
3. True parameter $\boldsymbol{\sigma}^2$ is bounded and $|\boldsymbol{\gamma}| leq 1$
4. True parameter $\boldsymbol{\beta}_n$ lies in interior of comapact subset $\mathcal{B} \in \mathbb{R}^{p_n}$
5. $\mathbb{X}$ is bounded element wise.
6. $0 < \lambda_{min}\big(\frac{1}{n}\sum_{i=1}^{n} \mathbb{X}_i^{\intercal}\mathbb{X}_i\big) \leq \lambda_{max}\big(\frac{1}{n}\sum_{i=1}^{n} \mathbb{X}_i^{\intercal}\mathbb{X}_i\big) < \infty$
7. $\min_{1\leq j\leq s_n} \boldsymbol{\beta}_{n0j}/\lambda_n \to \infty$ If $p_n = \mathbf{o}(n^{\alpha})$ then $\lambda_n = \mathbf{O}(n^{-\eta})$ where $0 < \alpha < \frac{4}{3}$ and $0 < \eta < (2-\alpha)$
8. $\max_i w_{i+_{,n}} = k < \infty$
9. $\sum_{i=1}^{s_n} |\boldsymbol{\beta}_{i,n}| < \infty$
10. $\min\limits_{1\leq i\leq p_n} \boldsymbol{\beta}_{n,0}(i)/\boldsymbol{\lambda} \to \infty$
11. $\frac{p_n^3}{n} = o(1)$
12. $\boldsymbol{\lambda}_n \to 0$
13. $p_n^2 log n^4 = o(n\boldsymbol{\lambda}_n^2)$

## 1.10 Consistency

**Theorem 1.10.1.** There exist approximate GEE solution of $n^{th}$ step be $\hat{\boldsymbol{\beta}}_{n,n}^{\intercal} = (\hat{\boldsymbol{\beta}}_{n1,n}^{\intercal}, \hat{\boldsymbol{\beta}}_{n2,n}^{\intercal})$

$$\mathbf{P}(|\mathbb{U}_{nj}\hat{\boldsymbol{\beta}})| = 0, \quad j = 1, 2, .....p_n) \to 1$$

$$\mathbf{P}(|\mathbb{U}_{nj}\hat{\boldsymbol{\beta}})| \leq \frac{\lambda_n}{logn}, \quad j = p_n + 1, p_n + 2, .....s_n) \to 1$$

$$\mathbf{P}(\hat{\boldsymbol{\beta}}_{n2} = \mathbf{0}) \to 1$$

**Remark 1.** The approximation of the solution is conveyed through second item which says that the value of elements from $p_n + 1$ to $s_n$ of score equation at $\hat{\boldsymbol{\beta}}_{n,n}$ are not exactly 0 but less than $\lambda_n/logn$ so by choosing $\lambda_n$ carefully we can attain good approximation

## 1.11 CLT

Define $\boldsymbol{\Xi}_n = \frac{\partial \boldsymbol{\mu}_{\intercal}}{\partial \boldsymbol{\beta}_n} V_n^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}_n}$

Define sandwich variance estimator by

$$\hat{H}_n = \hat{\boldsymbol{\Xi}}^{-1} \left( \frac{\partial \boldsymbol{\mu}_{\intercal}}{\partial \boldsymbol{\beta}_n} \hat{V}_n^{-1} (\boldsymbol{Y} - \hat{\boldsymbol{\mu}}(\boldsymbol{\beta}_n) (\boldsymbol{Y} - \hat{\boldsymbol{\mu}}(\boldsymbol{\beta}_n)^{\intercal} \hat{V}_n^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}_n} \right) \hat{\boldsymbol{\Xi}}^{-1}$$

$\forall \alpha_n \in \mathbb{R}^{p_n} \|\alpha_n\| = 1$

$$\alpha_n \, \hat{H}_n (\boldsymbol{\beta}_{n,0})^{-\frac{1}{2}} \frac{\partial \boldsymbol{\mu}_{\intercal}}{\partial \boldsymbol{\beta}_n} \hat{V}_n^{-1} (\boldsymbol{Y} - \hat{\boldsymbol{\mu}}(\boldsymbol{\beta}_{n,0}) (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) \longrightarrow \mathbb{N}(0, 1)$$

## 1.12 Simulation results

We simulate using $X = \mathbb{I}$ and maximal number of neighbors $k = 10$ . We record their std error and set $p = n^{1.5}$ and $s = n^{0.8}$ and $\sigma^2$=1

| $\boldsymbol{\sigma}^2$ | $n = 400, s_n = 30p_n$=**500** | | |
|---|---|---|---|
| | $\boldsymbol{\gamma} = 0.05$ | $\boldsymbol{\gamma} = 0.75$ | $\boldsymbol{\gamma} = 0.9$ |
| 0.1 | (0.28,0.81,0.85) | (0.27 ,0.73,0.76) | (0.38,0.85,.60) |
| 10 | (0.24, 0.74,0.84) | (0.21,0.73,0.85) | (0.35,0.82, 0.72) |
| 100 | (0.21,0.70,0.87) | (0.22,0.68, 0.86) | (0.32,0.83,0.85) |
| $\boldsymbol{\sigma}^2$ | $n = 1000, s_n = 30p_n = 2000$ | | |
| | $\boldsymbol{\gamma} = 0.05$ | $\boldsymbol{\gamma} = 0.75$ | $\boldsymbol{\gamma} = 0.9$ |
| 0.1 | (0.22,0.49,0.91) | (0.25,0.57, 0.89) | (0.31,0.61,0.83) |
| 10 | (0.21,0.27,0.91) | (0.23,0.59, 0.91) | (0.23,0.61,0.84) |
| 100 | (0.21,0.28,0.93) | (34.90 ,0.91) | (0.30,0.58,0.84) |

Table 1.1: Bias of absolute value,Sample Standard Deviation,Empirical Coverage probability of $\boldsymbol{\beta}$

# APPENDIX

# APPENDIX

**Variance Components**

$$\boldsymbol{V}_{3,n}(\boldsymbol{\beta}_n, \boldsymbol{\sigma}_n^2, \boldsymbol{\gamma}_n) = \mathbb{V}(\boldsymbol{Y}^*)$$

$$= \underbrace{\boldsymbol{V}_{\boldsymbol{\Psi}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\Psi}}(\boldsymbol{Y}^*)}_{\boldsymbol{V}_{1,n}} + \underbrace{\mathbb{E}_{\boldsymbol{\Psi}}\boldsymbol{V}_{\boldsymbol{Y}|\boldsymbol{\Psi}}(\boldsymbol{Y}^*)}_{\boldsymbol{V}_{2,n}}$$

$$\boldsymbol{V}_{1,n} = \boldsymbol{\sigma}^2(\mathbb{I} - \boldsymbol{\gamma}M_nW_n)^{-1}M_n$$

$$\boldsymbol{V}_{2,n} = Diag\left[\left(\frac{1}{E_i}exp\left(-\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2\right)\right) + \mathbb{E}_{\boldsymbol{\Psi}}\left(b_*^{'2}(\Psi_i)exp(\Psi_i) + b_*^2(\Psi_i)\right)\right]$$

(CRLB is attained for exponential family)

$$\boldsymbol{V}_{\boldsymbol{Y}|\boldsymbol{\Psi}}(\boldsymbol{Y}^*) = \left(\frac{\partial(\Psi + bias(\Psi))}{\partial(exp(E\Psi))}\right)^2 / I\big(exp(\psi)\big)$$

$$= exp(-\Psi/E)[1 + b_*(\Psi)]^2$$

Moment Generating function of Normal Distribution

$$\mathbb{E}_{\boldsymbol{\Psi}}(exp(-\Psi_i/E_i)) = exp\left(-\boldsymbol{X}_i^\intercal\boldsymbol{\beta} + \boldsymbol{\sigma}^2\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2\right)/E_i$$

Assuming negligible bias and derivative of bias of order $(\frac{1}{n})$

$$(\boldsymbol{V}_{2,n})_{ii} = \mathbb{E}_{\boldsymbol{\Psi}}\boldsymbol{V}_{\boldsymbol{Y}|\boldsymbol{\Psi}}(\boldsymbol{Y}_i^*) + \mathbb{E}_{\boldsymbol{\Psi}}\big(\boldsymbol{\Psi} - \boldsymbol{E}_{\boldsymbol{Y}|\boldsymbol{\Psi}}(\boldsymbol{Y}_i^*)\big)^\intercal\big(\boldsymbol{\Psi} - \boldsymbol{E}_{\boldsymbol{Y}|\boldsymbol{\Psi}}(\boldsymbol{Y}_i^*)\big)$$

$$= \mathbb{E}_{\boldsymbol{\Psi}}\boldsymbol{V}_{\boldsymbol{Y}|\boldsymbol{\Psi}}(\boldsymbol{Y}_i^*) + b_*^2(\Psi_i)$$

We will approximate to get

$$\boldsymbol{V}_{2,n} = Diag\Big( \frac{1}{E_i} exp\,(-\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) \Big)$$

**Inequalities using function $Y^*$**

*Proof.* We assume $e^{\Psi} = \lambda$ onwards Since $log(x+0.5)$ is a concave function. By Jensen's inequality $\mathbb{E}\, log(x+0.5) \leq log(\mathbb{E}\, x + 0.5) = log(\lambda + 0.5)$ and using $Y|\lambda \overset{ind}{\sim} Poisson(\lambda)$ so it follows Stein Chen identity Stein Chen identity, $\mathbb{E}(Yg(Y)) = \lambda \mathbb{E}(g(Y+1)$ when $\mathbb{E}|Yg(Y)|$ and $\mathbb{E}|g(Y+1)|$ exists. Hence Taylor approximation is valid and central moments can be derived using Stirling's number. From Log Sovolev inequality in Poisson Measure for any function $f : \mathbb{R} \to (0, \infty)$

$$\mathbb{E}_\lambda \Big[ f(Y) log\big(f(Y)\big) - \mathbb{E}[f(Y)]\, log\big(\,\mathbb{E}[f(Y)]\,\big) \Big] \leq \lambda\, \mathbb{E}\Big( \frac{\big(\,f(Y+1) - f(Y)\,\big)^2}{f(Y)} \Big)$$

Here $\mathbb{E}$ denotes $\mathbb{E}_\lambda$ Using $f(Y) = Y + \frac{1}{2}$ we obtain using Log sovolev inequality

$$\begin{aligned}
&\mathbb{E}\Big[ (Y+\frac{1}{2})\, \log\big(Y+\frac{1}{2}\big) - (\lambda+\frac{1}{2})\, log\big(\,\lambda+\frac{1}{2}\,\big) \Big] \\
&\leq \lambda\, \mathbb{E}\Big( \frac{1}{Y+\frac{1}{2}} \Big) \lambda\, \mathbb{E}\Big[ log\big(Y+\frac{1}{2}+1\big) - \log\big(\,\lambda+\frac{1}{2}\,\big) \Big] \;+\; \frac{1}{2}\, \mathbb{E}\Big[ log\big(Y+\frac{1}{2}\big) - \log\big(\,\lambda+\frac{1}{2}\,\big) \Big] \\
&\leq \lambda\, \mathbb{E}\Big( \frac{1}{Y+\frac{1}{2}} \Big) \qquad \lambda\, \mathbb{E}\Big[ log\big(Y+\frac{1}{2}\big) + \Big( \frac{1}{Y+\frac{1}{2}} \Big) -, \frac{1}{2} \Big( \frac{1}{Y+\frac{1}{2}+b} \Big)^2 - \log\big(\,\lambda+\frac{1}{2}\,\big) \Big] \\
&+ \frac{1}{2}\, \mathbb{E}\Big[ log\big(Y+\frac{1}{2}\big) - \log\big(\,\lambda+\frac{1}{2}\,\big) \Big] \leq \lambda\, \mathbb{E}\Big( \frac{1}{Y+\frac{1}{2}} \Big)
\end{aligned}$$

■

**Moments of Log Normal Poisson**

$$\mathbb{E}(\boldsymbol{Y}_i) = exp\left(\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + \boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2\right) = \mu_i$$

$$Var(\boldsymbol{Y}_i) = \mu_i + \mu_i^2 \left(exp(\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}) - 1\right)$$

$$Cov(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = \mu_i \mu_j \left(exp(\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ij}) - 1\right)$$

$$Corr(Y_i, Y_j) = \frac{exp(\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ij} - 1}{[exp(\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}) - 1 + \mu_i]^{\frac{1}{2}} \ [exp(\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{jj}) - 1 + \mu_j]^{\frac{1}{2}}}$$

$$\begin{aligned}\mathbb{E}\Big(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\Big)^3 &= \exp(3\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 9\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) - 3\exp(3\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 5\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) \\ &+2\exp(3\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 3\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) - 3\exp(2\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 4\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) \\ &-3\exp(2\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 2\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) + \exp(1\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 1\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2)\end{aligned}$$

$$\begin{aligned}\mathbb{E}(\boldsymbol{Y}_i - \mathbb{E}(\boldsymbol{Y}_i))^4 &= \exp(4\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 16\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) - 4\exp(4\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 10\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) \\ &+6\exp(4\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 6\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) - 3\exp(4\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 4\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) \\ &+6\exp(3\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 9\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) - 12\exp(3\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 5\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) \\ &+6\exp(3\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 3\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) + 7\exp(2\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 4\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) \\ &-4\exp(2\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 2\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2) + \exp(1\boldsymbol{X}_i^\intercal \boldsymbol{\beta} + 1\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2)\end{aligned}$$

**Lemma 1.12.1.**

$$Corr(Y_i, Y_j) \leq Corr(\Psi_i, \Psi_j)$$

*Proof.* Standardize $(\frac{\Psi_i}{\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}}, \frac{\Psi_j}{\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{jj}})$ Let resulting marginal model be $(Z_i, Z_j)$

Standardize $(\frac{Y_i}{Var(Y_i)}\,,\,\frac{Y_j}{Var(Y_j)})$

Use the inequality $Corr(Y_i,Y_j) = \frac{e^a-1}{\left(e^a-1+b\right)^{0.5}\left(e^a-1+c\right)^{0.5}} < a = Corr(Z_i,Z_j)$ for $|a|<1$ and $b>0\,,\,c>0$ ■

**Lemma 1.12.2.** Log normal distribution is dependent on first two moments of the latent model

*Proof.*

$$E(e^{it^\intercal \boldsymbol{Y}}) = E_{\boldsymbol{\Psi}} E_{\boldsymbol{Y}|\boldsymbol{\Psi}}(e^{it^\intercal \boldsymbol{Y}}) = E_{\boldsymbol{\Psi}}\big(\prod_{j=1}^{n} E_{\boldsymbol{Y}_j|\Psi_j}(e^{it_j^\intercal \boldsymbol{Y}})\big) = E_{\boldsymbol{\Psi}} \prod_{j=1}^{n} e^{\Psi_j(e^{it_j}-1)}$$

$$E_{\boldsymbol{\Psi}}\, e^{\sum_{j=1}^{n}\Psi_j\,(e^{it_j}-1)} = exp\Big((e^{it}-1)^\intercal \mathbb{X}\boldsymbol{\beta} + \frac{\boldsymbol{\sigma}^2\,\widetilde{(e^{it}-1)}^{\intercal}\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})\widetilde{(e^{it}-1)}}{2}\Big)$$

here $\widetilde{(e^{it}-1)}$ denotes vector $\widetilde{(e^{it}-1)}_j = (e^{it_j}-1)$ ■

**Relationship between induced and latent Model co variance**

$$\begin{aligned}(\boldsymbol{V}_{2,n})_{ii} &= \left(\frac{1}{E_i} exp\,(-\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2)\right)\\ &\leq \left(\frac{1}{E_i} exp\,(-\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\sigma}^2 \frac{1}{2(1-|\boldsymbol{\gamma}|)w_{i+}})\right)\\ &\geq \left(\frac{1}{E_i} exp\,(-\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\sigma}^2 \frac{1}{2w_{i+}})\right)\end{aligned}$$

**Range of $\boldsymbol{\gamma}$**

For the matrix $\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})$ to be positive definite matrix , it is sufficient for $\boldsymbol{\gamma} \in (-1,1)$ Since

$$\boldsymbol{V}_{1,n}^{-1}(\boldsymbol{\gamma}) = M_n^{-1} - \boldsymbol{\gamma} W_n$$

iff $|\boldsymbol{\gamma}| \leq 1$

Define strictly diagonal dominant matrix

$$\Delta(A)_i = |a_{ii}| - \sum_{i \neq j} |a_{ij}| \geq 0 \text{for all } i$$

*Proof.* Symmetric strictly diagonal dominant matrix with positive diagonal elements is positive definite from Varah (1975) .

$$\boldsymbol{V}_{1,n}^{-1}(\boldsymbol{\gamma}) = M_n^{-1} - \boldsymbol{\gamma} W_n$$

satisfies this condition for all $n$ iff $|\boldsymbol{\gamma}| \leq 1$

■

**Useful bounds about $\boldsymbol{V}_{1,n}$**

Define spectral norm $\|A\|_* = \max_i |\lambda_i(A)|$ Form the theory of diagonal dominant matrices we conclude following bounds. When $\boldsymbol{\gamma} > 0$ $\boldsymbol{V}_{1,n}^{-1}$ is M Matrix. We can exploit various results on M Matrix and extend them to the case $\boldsymbol{\gamma} < 0$ as $\Delta$ and $\|\|_\infty$ operator operates on absolute value of matrix entries.

1

$$\frac{1}{(1-|\boldsymbol{\gamma}|)\max_i wi+} \leq \| \left( \mathbb{I} - \boldsymbol{\gamma}\, \boldsymbol{M}_n \boldsymbol{W}_n \right)^{-1} M_n \|_* \leq \frac{1}{(1-|\boldsymbol{\gamma}|)\min_i wi+}$$

2

$$\frac{1}{w_{i+}} \leq \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii} \leq \frac{1}{(1-|\boldsymbol{\gamma}|)w_{i+}}$$

3

$$|\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ij}| \leq |\boldsymbol{\gamma}||\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}|$$

$$\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii} \geq \frac{1}{\boldsymbol{V}_{1,n}^{-1}(\boldsymbol{\gamma})_{ii}}$$

**Structure of $\boldsymbol{V}_{1,n}$** $\boldsymbol{V}_{1,n} = \left(\mathbb{I} - \boldsymbol{\gamma}\, \boldsymbol{M}_n \boldsymbol{W}_n\right)^{-1} M_n$ can be expanded by $\sum_{m=0}^{\infty} \boldsymbol{\gamma}^m \left(\boldsymbol{M}_n \boldsymbol{W}_n\right)^m M_n$ since

$\|(M_n W_n)\|_* < 1$ where $\|\|_*$ represents spectral norm or maximum of absolute eigenvalue.

**Lemma 1.12.3.** It can be proved by induction that $(M_n W_n)^r{}_{n\,i,j} \leq max_i \frac{1}{w_{i+}}$ for all $\geq 1$ and all $n$

*Proof.* Note that $(M_n W_n)_{i,m} = \frac{w_{i,m}}{w_{i+}}$ where $w_{i,m} = w_{m,i}$ is 0 or 1 $\sum_{j \in \mathcal{N}(i)} w_{ij} = w_{i+}$. Let $\boldsymbol{M}_{n\times n} =$ Diag $(\frac{1}{W_{i+}})_{ii}$

$$(M_n W_n)^1_{n\,i,m} = \frac{w_{i,m}}{w_{i+}} \leq max_i \frac{1}{w_{i+}}$$

$$(M_n W_n)^r_{n\,i,m} = \sum_{j=1}^{n} (M_n W_n)^1_{i,j} \, (M_n W_n)^{r-1}_{n\,j,m} \leq \sum_{j=0}^{n} (M_n W_n)_{i,j} \, max_i \frac{1}{w_{i+}} \leq \sum_{j=0}^{n} \frac{w_{i,j}}{w_{i+}} \, 1 \leq \, max_i \frac{1}{w_{i+}}$$

■

**Lemma 1.12.4.** $\|V_n^{-1}\|_\infty \geq \inf_i \boldsymbol{\mu}_i + \inf_i \boldsymbol{\mu}_i^2 \frac{\boldsymbol{\sigma}^2}{w_{i+}(1-\boldsymbol{\gamma})}$

*Proof.* Case $\boldsymbol{\gamma} > 0$:

Here all entries of matrix $(M_n^{-1} - \boldsymbol{\gamma} W_n)^{-1} = \boldsymbol{V}_{1,n}$ are positive. The row sum can be written as

$$\boldsymbol{V}_{1,n}^{-1} \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix} = (1-\boldsymbol{\gamma}) \begin{pmatrix} w_{1+} \\ w_{2+} \\ \cdots \\ w_{i+} \\ \cdots \\ w_{n+} \end{pmatrix} \leq \max_i w_{i+} \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix}$$

$$\left\{exp(\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n\ i,j}) - 1\right\} \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix} \geq \boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n} \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix} \geq \frac{\boldsymbol{\sigma}^2}{\max_i \ w_{i+}(1-\boldsymbol{\gamma})} \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix}$$

$$\|V_n\|_\infty \geq \inf_i \boldsymbol{\mu}_i + \inf_i \boldsymbol{\mu}_i^2 \frac{\boldsymbol{\sigma}^2}{w_{i+}(1-\boldsymbol{\gamma})}$$

■

**Lemma 1.12.5.** If $1 - \boldsymbol{\gamma} - \frac{\boldsymbol{\gamma}^2}{1-\boldsymbol{\gamma}^2}(1 - 2\frac{1}{k}) > 0$ is strictly diagonally dominant matrix

*Proof.* Since $(\boldsymbol{M}_n \boldsymbol{W}_n)$ is non negative irreducible matrix by Perron Frobenius therom: for all $k$

$$(\boldsymbol{M}_n \boldsymbol{W}_n)^k \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix} = 1 \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix}$$

. using this result geometric expansion of $\boldsymbol{V}_{1,n} = \sum_{m=0}^{\infty} \boldsymbol{\gamma}^m \ (\boldsymbol{M}_n \boldsymbol{W}_n)^m M_n$ we can arrive at necessary and sufficient condition $1 - \boldsymbol{\gamma} + \min_i \sum_{m=2}^{\infty} \boldsymbol{\gamma}^m \left(2(\boldsymbol{M}_n \boldsymbol{W}_n)_{i,i}^m - 1\right) > 0$ for strict diagonal dominance of $\boldsymbol{V}_{1,n}$ Consider wort scenario where $(\boldsymbol{M}_n \boldsymbol{W}_n)_{i,i}^{2k+1} = 0$ then above condition is equivalent to $1 - \boldsymbol{\gamma} - \frac{\boldsymbol{\gamma}^2}{1-\boldsymbol{\gamma}^2}(1 - 2\frac{1}{k}) > 0$ which is cubic in gamma. ■

**Lemma 1.12.6.** If $\boldsymbol{V}_{1,n}$ is strictly diagonally dominant matrix then matrix $(V_n)_{i,j} = f(\boldsymbol{V}_{1,n})_{i,j}$ is strictly diagonally dominant matrix or equivalently if

$$|V_{1,n\,i,i}| \geq \sum_{j \neq i} |V_{1,n\,i,j}|$$

then

$$|f(V_{1,n\,i,i})| \geq \sum_{j \neq i} |f(V_{1,n\,i,j})|$$

where $f(x) = exp(\boldsymbol{\sigma}^2\, x) - 1$

*Proof.* We need to show that if from the results 1.12 it is known that $C_{W,\boldsymbol{\gamma}} > V_{n\,i,i} > 0$ and $V_{n\,i,i} > |\boldsymbol{\gamma}||V_{n\,i,j}|$ for all $i \neq j$. Using expansion

$$exp((x) - 1 = \sum_{i=0}^{\infty} \frac{x^k}{K!}$$

We prove that for all n and each $k$ by induction

$$V_{n,i,i}^k = V_{n,i,i}^{k-1} V_{n,i,i} \geq V_{n,i,i}^{k-1} \sum_{j \neq i} |V_{1,n\,i,j}|$$

$$\geq \max_{j \neq i} |V_{n,i,j}^{k-1}| \sum_{j \neq i} |V_{1,n\,i,j}| \geq \sum_{j \neq i} |V_{1,n\,i,j}^k|$$

■

The above result can hold large class smooth function $f : \mathbb{R} \to \mathbb{R}$ with domain being variance matrix elements.

**Lemma 1.12.7.** $\|V_n^{-1}\|_* \geq \inf_i \frac{w_{i+}(1+|\boldsymbol{\gamma}|)}{w_{i+}(1+|\boldsymbol{\gamma}|)\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^2 \boldsymbol{\sigma}^2}$

*Proof.*

$$\|V_n\|_* \leq \|V_n\|_\infty$$
$$\min_i \lambda_i(V_n^{-1}) \geq \frac{1}{\|V_n\|_\infty}$$

■

**Lemma 1.12.8.** If $A, B, C, A - B$ be symmetric positive semidefinite matrix then

$$\min_i \lambda(i)(A + C) \geq \min_i \lambda(i)(B + C)$$

*Proof.* Using results in Hiai & Lin (2017) take $k = n$ to obtain

$$\min_i \lambda(i)(AC^{-1}) \geq \min_i \lambda(i)(BC^{-1})$$
$$\min_i \lambda(i)(AC^{-1} + \mathbb{I}) \geq \min_i \lambda(i)(BC^{-1} + \mathbb{I}) \qquad \text{Since all matrices commute with } \mathbb{I}$$
$$\min_i \lambda(i)\big((AC^{-1} + \mathbb{I})C\big) \geq \min_i \lambda(i)\big((BC^{-1} + \mathbb{I})C\big)$$
$$\min_i \lambda(i)(A + C) \geq \min_i \lambda(i)(B + C)$$

■

**Lemma 1.12.9.**

$$\|V_n^{-1}\|_* \leq \frac{1}{\exp[\min_i \lambda(i)(V_{1,n})] - 1} \leq \frac{1}{\exp[\frac{1}{(1-\boldsymbol{\gamma})\min_i w+_i}] - 1}$$

*Proof.* Since

$$V_n = exp((V_{1,n}) - 1 = \sum_{i=0}^{\infty} \frac{V_{1,n}^{\circ k}}{K!}$$

where $\circ$ represents hadamard power We can obtain trivial inequality

$$\lambda_i(V_n) \geq \lambda_i(V_{1,n})$$

But we can obtain bounds dependent upon function form $f$

We use results

$$\min_i \lambda(i)(A \circ B) \geq \min_i \lambda(i)(A) \min_i \lambda(i)(B)$$

and using induction and 1.12.8 we conclude that

$$\min_i \lambda(i)[exp(V_{1,n}) - 1] \geq \exp[\min_i \lambda(i)(V_{1,n})] - 1$$

■

**Lemma 1.12.10.** $\|V_n\|_\infty \geq \sup_i(\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^2\boldsymbol{\sigma}^2 \frac{1}{w_{i+}(1+|\boldsymbol{\gamma}|)}\ )$

*Proof.*

$$V_n(\boldsymbol{\sigma}^2, \boldsymbol{\gamma}, \boldsymbol{\beta}_n) = Diag(\boldsymbol{\mu}_i)\ +\ Diag(\boldsymbol{\mu}_i)[\boldsymbol{R}^{\odot 2}(\boldsymbol{\sigma}^2, \boldsymbol{\gamma}) - \boldsymbol{J}_n]Diag(\boldsymbol{\mu}_i)$$

$$\left(\boldsymbol{R}^{\odot 2}(\boldsymbol{\sigma}^2, \boldsymbol{\gamma}) - \boldsymbol{J}_n\right)_{i,j} = exp(\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n\ \ i,j}) - 1$$

Using the expansion to obtain $\frac{\boldsymbol{\sigma}^2}{\max_{i,j} \boldsymbol{V}_{1,n\ \ i,j}} = \tilde{\sigma}^2$

$$\exp(\tilde{\sigma}^2 x) - 1 - \sigma^2 x = \sigma^2\left(x^2/2! + x^3/3! + ... + x^2 n/2n! + x^{2n+1}/(2n+1)!\right) \geq 0$$

$$\exp(\tilde{\sigma}^2 x) - 1 \geq \tilde{\sigma}^2 x$$

for all $x \in [-1,1]$ Therefore

$$\|V_n\|_\infty \geq \sup_i(\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^2\|V_{1,n}\|_\infty) \geq \sup_i(\boldsymbol{\mu}_i + \frac{\boldsymbol{\mu}_i^2\sigma^2}{w_{i+}(1+|\boldsymbol{\gamma}|)}\ )$$

■

**Lemma 1.12.11.** $\|V_n\|_\infty \leq \sup_i \frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_i^2\left(exp(\boldsymbol{\sigma}^2\boldsymbol{V}_{1,n\ \ i,i})-1\right)}{\boldsymbol{\sigma}^2\boldsymbol{V}_{1,n\ \ i,i}}\ \frac{1}{w_{i+}(1-|\boldsymbol{\gamma}|)}$

*Proof.* We use the previous result $Corr(Y_i,Y_j) \leq Corr(\Psi_i,\Psi_j)$ for all $i,j$ hence

$$\|Diag\left(var^{-1}(Y_i)\right)V_n\,Diag\left(var^{-1}(Y_i)\right)\|_\infty \leq \|Diag\left(var^{-1}(\Psi_i)\right)V_{1,n}\,Diag\left(var^{-1}(\Psi_i)\right)\|_\infty$$

$\|V_n\|_\infty \leq \sup_i \frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_i^2\left(exp(\boldsymbol{\sigma}^2\boldsymbol{V}_{1,n\ \ i,i})-1\right)}{\boldsymbol{V}_{1,n\ \ i,i}}\ \frac{1}{w_{i+}(1-|\boldsymbol{\gamma}|)}$ ■

**Derivation of estimating equations**

1 derivation of $\mathbb{U}_n$ is straightforward.

2

$$\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\gamma}} =$$

$$\frac{\partial\,\mathbf{exp}\left(-\boldsymbol{X}^\intercal\boldsymbol{\beta}_\mathbf{n} + \boldsymbol{\sigma}^\mathbf{2}\,\mathbf{Diag}(\boldsymbol{V}_{\mathbf{1,n}}(\boldsymbol{\gamma}))_{\mathbf{n}\times\mathbf{1}}/\mathbf{2}\right)}{\partial\boldsymbol{\gamma}} =$$

$$\boldsymbol{\sigma}^2/2\begin{pmatrix}\boldsymbol{\mu}_1 & & \\ & \ddots & \\ & & \boldsymbol{\mu}_n\end{pmatrix} Diag(\frac{\partial\boldsymbol{V}_{\mathbf{1,n}}(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}) =$$

$$\boldsymbol{\sigma}^2/2\begin{pmatrix}\boldsymbol{\mu}_1(\boldsymbol{V}_{1,n}^{-1}\boldsymbol{W}_n\boldsymbol{V}_{1,n}^{-1})_{1,1} & & \\ & \ddots & \\ & & \boldsymbol{\mu}_n(\boldsymbol{V}_{1,n}^{-1}\boldsymbol{W}_n\boldsymbol{V}_{1,n}^{-1})_{n,n}\end{pmatrix} =$$

$$\boldsymbol{\sigma}^2/2 \; Diag(\boldsymbol{V}_{\mathbf{1,n}}^{-\mathbf{1}}\boldsymbol{W}_{\mathbf{n}}\boldsymbol{V}_{\mathbf{1,n}}^{-\mathbf{1}}) \; \boldsymbol{\mu}$$

3 Derivation of $\mathbb{F}'_{2,n}$ is also similar

**Stationary Case** Joint estimate of parameters may be difficult to compute. We solve for simplest case first where

$$(W_n)_{i,j} = f(|i-j|)\mathbb{I}(|i-j| \leq k)$$

for some fixed k. Therefore yielding matrix $\boldsymbol{V}_{1,n}(\boldsymbol{\gamma}) = (M_n^{-1} - \boldsymbol{\gamma}W_n)^{-1}$ stationary co variance matrix. It can be seen by Cramer's rule $(\boldsymbol{V}_{1,n})_{i,j} = \frac{\det[(\boldsymbol{V}_{1,n}^{-1})_{-(i,j)}]}{\det[(\boldsymbol{V}_{1,n})^{-1}]}$

**Re parametrization**

$$\begin{aligned}
&\boldsymbol{\mu}_i(\boldsymbol{\beta}_n, \boldsymbol{\sigma}_2, \boldsymbol{\gamma}) = exp\left(-\boldsymbol{X}_i^\intercal \boldsymbol{\beta}_n + \boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2\right) \\
&= \exp(-[\boldsymbol{X}_i^\intercal, 1] \begin{bmatrix} \boldsymbol{\beta}_n \\ \boldsymbol{\sigma}^2 \frac{V_0}{2} \end{bmatrix}) = \exp(-[\boldsymbol{X}_i^\intercal, 1] \begin{bmatrix} \boldsymbol{\beta}_n \\ \tilde{\boldsymbol{\sigma}}^2 \end{bmatrix}) \\
&= \exp(-\tilde{\boldsymbol{X}}_i^\intercal \tilde{\boldsymbol{\beta}}) = \boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}_n})
\end{aligned}$$

since

$\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii} = V_0 \;_{1\times 1}$ is same for all i due to stationarity

$$V_n(\boldsymbol{\sigma}^2, \boldsymbol{\gamma}, \boldsymbol{\beta}_n) = Diag(\boldsymbol{\mu}_i) \; + \; Diag(\boldsymbol{\mu}_i)^\intercal [\boldsymbol{R}^{\odot 2}(\tilde{\boldsymbol{\sigma}}^2, \boldsymbol{\gamma}) - \boldsymbol{J}_n] Diag(\boldsymbol{\mu}_i)$$

$$\left(\boldsymbol{R}^{\odot 2}(\tilde{\boldsymbol{\sigma}}^2, \boldsymbol{\gamma}) - \boldsymbol{J}_n\right)_{i,j} = exp\left(\tilde{\boldsymbol{\sigma}}^2 \frac{\boldsymbol{V}_{1,n \;\; i,j}}{V_0}\right) - 1$$

Here $\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii} = V_0\ _{1\times 1}$ is same for all i due to stationarity

**GEE in stationary setup**

1 Score equation for mean

$$\mathbb{U}_n\big(\tilde{\boldsymbol{\beta}}_n\big)_{(p_n+1)\times 1} = \underbrace{\frac{1}{n}\frac{\partial \boldsymbol{\mu}}{\partial \tilde{\boldsymbol{\beta}}}^{\intercal}\big[\mathbf{V_n}(\tilde{\boldsymbol{\sigma}}^{\mathbf{2}}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}_{\mathbf{n}})\big]^{-1}\Big(\boldsymbol{Y} - \boldsymbol{\mu}(\tilde{\boldsymbol{\beta}}_n)\Big)}_{\tilde{S}_n} + \begin{bmatrix} \mathbf{q}(\boldsymbol{\lambda}_n)(|\tilde{\boldsymbol{\beta}}_{n\ ,-n}|) \odot sgn(\tilde{\boldsymbol{\beta}}_{n\ ,-n}) \\ 0 \end{bmatrix}$$

under $\boldsymbol{\lambda}_n \geq 0$ and $\tilde{\boldsymbol{\beta}}_{n\ ,-n\,p_n\times 1}$ denote sub vector such that

$$\begin{bmatrix} \tilde{\boldsymbol{\beta}}_{n\ ,-n} \\ \tilde{\boldsymbol{\beta}}_{n\ ,n} \end{bmatrix} = \tilde{\boldsymbol{\beta}}_n$$

$\mathbf{q}(\boldsymbol{\lambda}_n)$ is derivative of SCAD penalty function wrt $\boldsymbol{\beta}$ with penalty parameter $\boldsymbol{\lambda}_n$

2 Method of Moment

$$\exp\left(\tilde{\boldsymbol{\sigma}}^2\ \frac{V_t(\boldsymbol{\gamma})}{V_0(\boldsymbol{\gamma})}\right) - 1 = \mathbb{E}\Big[\frac{\sum_{i=t+1}^{n}\Big(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}}_n)\Big)\Big(\boldsymbol{Y}_{i+t} - \boldsymbol{\mu}_{i+t}(\tilde{\boldsymbol{\beta}}_n)\Big)}{\sum_{i=t+1}^{n}\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}}_n)\boldsymbol{\mu}_{i+t}(\tilde{\boldsymbol{\beta}}_n)}\Big]$$

2.1 Score for scale parameter

$$\hat{\exp}(\tilde{\boldsymbol{\sigma}}^2) = \frac{\sum_{i=1}^{n}\Big(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i(\tilde{\boldsymbol{\beta}}_n)\Big)^2 - \hat{\boldsymbol{\mu}}_i(\tilde{\boldsymbol{\beta}}_n)}{\sum_{i=1}^{n}\hat{\boldsymbol{\mu}}_i(\tilde{\boldsymbol{\beta}}_n)^2} + 1$$

$$\hat{\tilde{\boldsymbol{\sigma}}}^2 = \log\Big[\frac{\sum_{i=1}^{n}\Big(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i(\tilde{\boldsymbol{\beta}}_n)\Big)^2 - \hat{\boldsymbol{\mu}}_i(\tilde{\boldsymbol{\beta}}_n)}{\sum_{i=1}^{n}\hat{\boldsymbol{\mu}}_i(\tilde{\boldsymbol{\beta}}_n)^2} + 1\Big]$$

$\hat{\tilde{\boldsymbol{\sigma}}}^2$ can be negative by using this method.

2.2 Score for correlation parameter

$$\frac{\hat{V}_t(\boldsymbol{\gamma})}{\hat{V}_0(\boldsymbol{\gamma})} = \log\left[\frac{\sum_{i=t+1}^{n}\left(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i(\tilde{\boldsymbol{\beta}}_n)\right)\left(\boldsymbol{Y}_{i+t} - \hat{\boldsymbol{\mu}_{i+t}}(\tilde{\boldsymbol{\beta}}_n)\right)}{\sum_{i=t+1}^{n}\hat{\boldsymbol{\mu}}_i(\tilde{\boldsymbol{\beta}}_n)\hat{\boldsymbol{\mu}_{i+t}}(\tilde{\boldsymbol{\beta}}_n)} + 1\right]/\hat{\tilde{\boldsymbol{\sigma}}}^2$$

$\hat{\boldsymbol{\gamma}}$ can be outside (-1,1)

$V_t(\boldsymbol{\gamma}) = V_{|i-j|=t}(\boldsymbol{\gamma})$ due to stationarity time series of counts.

3 Cressie variogram estimate

$$\mathbb{E}\left[\frac{\sum_{i=t+1}^{n}(Y_i - \mathbb{E}Y_i)(Y_i - \mathbb{E}Y_{i+t})}{n-t}\right] = VAR(Y_i) + VAR(Y_{i+t}) - 2COV(Y_i, Y_{i+t})$$

Thus we obtain $\mathbb{E}\left[\frac{\sum_{i=t+1}^{n}\left(\frac{\boldsymbol{Y}_i}{\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}}_n)} - \frac{\boldsymbol{Y}_{i+t}}{\boldsymbol{\mu}_{i+t}(\tilde{\boldsymbol{\beta}}_n)}\right)^2 - \hat{\boldsymbol{\mu}}_i(\tilde{\boldsymbol{\beta}}_n)^{\frac{-1}{2}} - -\hat{\boldsymbol{\mu}}_{i+t}(\tilde{\boldsymbol{\beta}}_n)^{\frac{-1}{2}}}{2(n-t)}\right] =$

$\exp(\tilde{\boldsymbol{\sigma}}^2) - \exp(\tilde{\boldsymbol{\sigma}}^2\, V_t(\boldsymbol{\gamma}))$

We take $t = 1$ find $\hat{\boldsymbol{\gamma}}$ form $\hat{V}_1(\boldsymbol{\gamma})$ when $V_1(\boldsymbol{\gamma})$ is known function of $\boldsymbol{\gamma}$.

Cressie showed that when dimension of $\boldsymbol{\beta}$ is fixed given $\boldsymbol{\sigma}^2$ $|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}| = O_p(\frac{1}{\sqrt{n}})$

**Matrix Inverse approximation** $\boldsymbol{V}_{1,n}$ is diagonal matrix

$$(\boldsymbol{V}_{3,n})^{-1}$$
$$= (\boldsymbol{V}_{1,n} + \boldsymbol{\sigma}^2\ \boldsymbol{V}_{2,n})^{-1}$$
$$= (\boldsymbol{V}_{2,n})^{-1} - (\boldsymbol{\sigma}^2\ \boldsymbol{V}_{2,n})^{-1}\left((\boldsymbol{V}_{1,n})^{-1} + (\boldsymbol{\sigma}^2\ \boldsymbol{V}_{2,n})^{-1}\right)^{-1}(\boldsymbol{\sigma}^2\ \boldsymbol{V}_{2,n})^{-1}$$
$$= (\boldsymbol{\sigma}^2\ \boldsymbol{V}_{2,n})^{-1} - (\boldsymbol{\sigma}^2\ \boldsymbol{V}_{2,n})^{-1}\left(\boldsymbol{M}_n^{-1} + (\boldsymbol{\sigma}^2\ \boldsymbol{V}_{2,n})^{-1} - \boldsymbol{\gamma}\boldsymbol{W}_n\right)^{-1}(\boldsymbol{\sigma}^2\ \boldsymbol{V}_{2,n})^{-1}$$
$$\boldsymbol{\sigma}^2\ \boldsymbol{V}_{2,n})^{-1} - (\boldsymbol{\sigma}^2\ \boldsymbol{V}_{2,n})^{-1}(\boldsymbol{\sigma}^2\ \boldsymbol{V}_{2,n} + \boldsymbol{M}_n^{-1})^{\frac{-1}{2}}$$
$$\left(\mathbb{I} - \boldsymbol{\gamma}(\boldsymbol{\sigma}^2\ \boldsymbol{V}_{2,n} + \boldsymbol{M}_n^{-1})^{\frac{-1}{2}}\boldsymbol{W}_n\,\boldsymbol{\sigma}^2\ \boldsymbol{V}_{2,n} + \boldsymbol{M}_n^{-1})^{\frac{-1}{2}}\right)^{-1}$$

$$( \boldsymbol{\sigma}^2 \boldsymbol{V}_{2,n} + \boldsymbol{M}_n^{-1})^{\frac{-1}{2}} ( \boldsymbol{\sigma}^2 \boldsymbol{V}_{2,n})^{-1}$$

**Lemma 1.12.12.** The spectral norm of

$( \boldsymbol{\sigma}^2 \boldsymbol{V}_{2,n} + \boldsymbol{M}_n^{-1})^{\frac{-1}{2}} \boldsymbol{W}_n ( \boldsymbol{\sigma}^2 \boldsymbol{V}_{2,n} + \boldsymbol{M}_n^{-1})^{\frac{-1}{2}}$

is same as spectral norm $( \boldsymbol{\sigma}^2 \boldsymbol{V}_{2,n} + \boldsymbol{M}_n^{-1})^{-1} W_n$ which is less than max row sum of

$( \boldsymbol{\sigma}^2 \boldsymbol{V}_{2,n} + \boldsymbol{M}_n^{-1})^{-1} W_n$

Hence the spectral norm of $( \boldsymbol{\sigma}^2 \boldsymbol{V}_{2,n} + \boldsymbol{M}_n^{-1})^{\frac{-1}{2}} \boldsymbol{W}_n ( \boldsymbol{\sigma}^2 \boldsymbol{V}_{2,n} + \boldsymbol{M}_n^{-1})^{\frac{-1}{2}} \leq 1$

$\left( \mathbb{I} - \boldsymbol{\gamma} ( \boldsymbol{\sigma}^2 \boldsymbol{V}_{2,n} + \boldsymbol{M}_n^{-1})^{\frac{-1}{2}} \boldsymbol{W}_n ( \boldsymbol{\sigma}^2 \boldsymbol{V}_{2,n} + \boldsymbol{M}_n^{-1})^{\frac{-1}{2}} \right)$ can be approximated by

$$\sum_{m=0}^{n} \boldsymbol{\gamma}^m \big( ( \boldsymbol{\sigma}^2 \boldsymbol{V}_{2,n} + \boldsymbol{M}_n^{-1})^{\frac{-1}{2}} \boldsymbol{W}_n ( \boldsymbol{\sigma}^2 \boldsymbol{V}_{2,n} + \boldsymbol{M}_n^{-1})^{\frac{-1}{2}} \big)^m + \mathbb{O}(\boldsymbol{\gamma}^n)$$

**Remark 2.** Hence forward $\boldsymbol{Y}^*$ will be denoted as $\boldsymbol{Y}$

**Bound on Frobenius norm of score**

Further

$$\mathbb{E} \left\| \mathbb{X}^\intercal \big[ \mathbf{V_{3,n}}(\boldsymbol{\gamma}) \big]^{-1} (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}) \right\|_2^2 \leq \mathbb{E} \left\| \mathbb{X}^\intercal \mathbf{D_n}(\boldsymbol{\gamma})^{-1} (\boldsymbol{Z} - \mathbb{X}\boldsymbol{\beta}) \right\|_2^2$$

where is $D_n$ is diagonal positive definite matrix.So $\boldsymbol{Z}$ have mutually independent random elements.

Using

$$\text{Since } (\boldsymbol{V}_{2,n})^{-1} \left( (\boldsymbol{V}_{1,n})^{-1} + (\boldsymbol{V}_{2,n})^{-1} \right)^{-1} (\boldsymbol{V}_{2,n})^{-1} \text{is Postive definite Matrix}$$

$$\mathbb{E} \left\| \mathbb{X}^\intercal \big[ \mathbf{V_{3,n}}(\boldsymbol{\gamma}) \big]^{-1} (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}) \right\|_2^2$$
$$= \mathbb{TR}(X^\intercal \boldsymbol{V}_{3,n}^{-1} X)$$

$$\leq \lambda_{max}(\boldsymbol{V}_{3,n}^{-1})tr(X^{\intercal}X)$$

$$\leq \lambda_{max}(\boldsymbol{V}_{2,n}^{-1})tr(X^{\intercal}X)$$

$$\leq \mathbb{E}\left\|\mathbb{X}^{\intercal}\left[\mathbf{D_n}(\boldsymbol{\gamma})\right]^{-1}(\boldsymbol{Z}-\mathbb{X}\boldsymbol{\beta})\right\|_2^2$$

where $\mathbf{D_n} = \frac{1}{\min\limits_i E_i} exp\left(-\boldsymbol{X}_i^{\intercal}\boldsymbol{\beta}\right) - \frac{\boldsymbol{\sigma}^2}{2}\frac{1}{\max\limits_i w_{i+}}$

Since $\lambda_{max}(\boldsymbol{V}_{2,n}^{-1}) \leq \left(\frac{1}{\min\limits_i E_i} exp\left(-\boldsymbol{X}_i^{\intercal}\boldsymbol{\beta}\right) - \boldsymbol{\sigma}^2\frac{1}{2\max\limits_i w_{i+}}\right)$

**Convergence of scale and correlation parameters**

$$|\hat{\boldsymbol{\sigma}^2}_n - \boldsymbol{\sigma}^2| = \mathbb{O}_p(\|\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n\|_2)$$

$$|(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma})| = \mathbb{O}_p(\|\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n\|_2)$$

**Convergence of variance parameter** In Yasui and Lele [2012] estimation of $\sigma^2$ are done using both hierarchical and marginal methods of conditional least squares

Hierarchical method

$$\hat{\boldsymbol{\sigma}^2}_n = \frac{1}{n}(\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n})^{\intercal}(\boldsymbol{M}_n^{-1} - \hat{\boldsymbol{\gamma}}_n\boldsymbol{W}_n)(\boldsymbol{Y}^* - \quad \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n}) - \frac{1}{n}(\boldsymbol{Y}^{**} - \boldsymbol{Y}^*)^{\intercal}M_n^{-1}(\boldsymbol{Y}^{**} - \boldsymbol{Y}^*)$$

Marginal method

$$\hat{\boldsymbol{\sigma}^2}_n = \frac{1}{n}(\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n})^{\intercal}(\boldsymbol{M}_n^{-1} - \hat{\boldsymbol{\gamma}}_n\boldsymbol{W}_n)(\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n}) - \frac{1}{n}\ \mathbb{TR}\left(\ \hat{\boldsymbol{V}_{2,n-1}}(\hat{\boldsymbol{\beta}}_{n,n}, \hat{\boldsymbol{\sigma}^2}_{n-1}, \hat{\boldsymbol{\gamma}}_n)\boldsymbol{M}_n^{-1}\right) \tag{1.1}$$

For both of these methods, it can be shown that rates are same

*Proof.*

$$
\begin{aligned}
&\hat{\boldsymbol{\sigma}^2}_n - \boldsymbol{\sigma}^2 \\
&= \frac{1}{n}(\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n})^\intercal(\boldsymbol{M}_n^{-1} - \hat{\boldsymbol{\gamma}}_n\boldsymbol{W}_n)(\boldsymbol{Y}^* \\
&- \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n}) - \frac{1}{n}(\boldsymbol{Y}^* - \mathbb{X}\boldsymbol{\beta}_n)^\intercal(\boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}_n\boldsymbol{W}_n)(\boldsymbol{Y}^* - \mathbb{X}\boldsymbol{\beta}_n) \\
&- \frac{1}{n}\mathbb{TR}\bigg(\hat{\boldsymbol{V}_{2,n-1}}(\hat{\boldsymbol{\beta}}_{n,n}, \hat{\boldsymbol{\sigma}^2}_{n-1}, \hat{\boldsymbol{\gamma}}_n)\boldsymbol{M}_n^{-1})\bigg) + \frac{1}{n}\mathbb{TR}\bigg(\boldsymbol{V}_{2,n}(\boldsymbol{\beta}_n, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})\boldsymbol{M}_n^{-1})\bigg) \\
&= \frac{1}{n}(\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n})^\intercal\big(\ (\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma})\boldsymbol{W}_n\ \big)(\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n}) \\
&+ \frac{1}{n}2\ (\boldsymbol{Y}^{*\intercal} - \mathbb{X}\boldsymbol{\beta}_n)(\boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\boldsymbol{W}_n)\mathbb{X}(\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n) + \frac{1}{n}(\hat{\boldsymbol{\beta}}_{n,n} \\
&- \boldsymbol{\beta}_n)^\intercal\mathbb{X}^\intercal(\boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\boldsymbol{W}_n)\mathbb{X}(\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n) \\
&+ \frac{1}{n}\ \mathbb{TR}\bigg(\big(\ \hat{\boldsymbol{V}_{2,n-1}}(\hat{\boldsymbol{\beta}}_{n,n}, \hat{\boldsymbol{\sigma}^2}_{n-1}, \hat{\boldsymbol{\gamma}}_n) - \boldsymbol{V}_{2,n}(\boldsymbol{\beta}_n, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})\ \big)\boldsymbol{M}_n^{-1}\bigg)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\bigg(\|(\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n})^\intercal\big(\ (\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma})\boldsymbol{W}_n\ \big)(\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n})\|_2\bigg) \\
&\leq \mathbb{E}\|(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma})\|_2\ \ \mathbb{E}\|(\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n})^\intercal\big(\ \boldsymbol{W}_n\ \big)(\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n})\|_2 \\
&= \|(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma})\|_2\mathbb{TR}(\ (\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n})^\intercal\big(\ \boldsymbol{W}_n\ \big)(\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n})) \\
&\leq\ n\ \|(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma})\|_2\ \lambda_{max}(\boldsymbol{V}_{3,n}\boldsymbol{W}_n) \\
&\leq\ n\ \|(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma})\|_2\ \ \max_i(\boldsymbol{V}_{3,n})_{ii}\lambda_{max}(\boldsymbol{W}_n)
\end{aligned}
$$

$$
\begin{aligned}
&\|\ (\boldsymbol{Y}^{*\intercal} - \mathbb{X}\boldsymbol{\beta}_n)(\boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\boldsymbol{W}_n)\mathbb{X}(\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n)\|_2 \\
&\leq \|\ (\boldsymbol{Y}^{*\intercal} - \mathbb{X}\boldsymbol{\beta}_n)(\boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\boldsymbol{W}_n)\mathbb{X}\|_2\ \ \|(\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n)\|_2 \\
&\leq\ \ max_i\ \boldsymbol{\lambda}_i\bigg((\boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\boldsymbol{W}_n)\boldsymbol{V}_3(\boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\boldsymbol{W}_n)\bigg)\ \sqrt{\mathbb{TR}(\mathbb{X}^\intercal\mathbb{X})}\ \ \|(\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n)\|_2
\end{aligned}
$$

$$= \|(\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n)\|_2^2 \ max_i \ \boldsymbol{\lambda}_i\Big((\boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\boldsymbol{W}_n)\boldsymbol{V}_3(\boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\boldsymbol{W}_n)\Big) \ \sqrt{\mathbb{TR}(\mathbb{X}\mathbb{X}^\intercal)}$$

$$= \mathbb{O}_p\sqrt{(\frac{p_n}{n})} \ \mathbb{O}\sqrt{(n\,p_n)}$$

$$= \mathbb{O}_p(p_n)$$

$$(\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n)^\intercal\mathbb{X}^\intercal(\boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\boldsymbol{W}_n)\mathbb{X}(\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n)$$

$$\leq \ \|(\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n)\|_2^2 \ max_i\boldsymbol{\lambda}_i \ \ (\mathbb{X}^\intercal(\boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\boldsymbol{W}_n) \ \ \mathbb{X})$$

$$\leq \mathbb{O}_p \ \ n \ \ (\frac{p_n}{n})$$

$$= \mathbb{O}_p \ (p_n)$$

By recursion when $n$ is large $\hat{\boldsymbol{\sigma}}_{n-1}^2 - \boldsymbol{\sigma}^2$ is same order of $\hat{\boldsymbol{\sigma}}_n^2 - \boldsymbol{\sigma}^2$

$$\boldsymbol{V}_{2,n}(\boldsymbol{\beta}_n, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})$$

$$= Diag\Big(\ \frac{1}{E_i}exp\,(-\boldsymbol{X}_i\boldsymbol{\beta} + \frac{\boldsymbol{\sigma}^2}{2}\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii})\Big)$$

$$\mathbb{TR}\big(\ \hat{\boldsymbol{V}_{2,n-1}}(\hat{\boldsymbol{\beta}}_{n,n}, \hat{\boldsymbol{\sigma}^2}_{n-1}, \hat{\boldsymbol{\gamma}}_n)\boldsymbol{M}_n^{-1}$$

$$- \boldsymbol{V}_{2,n}(\boldsymbol{\beta}_n, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}) \ \boldsymbol{M}_n^{-1}\big)$$

$$\leq \mathbb{TR}\big(\ \hat{\boldsymbol{V}_{2,n-1}}(\hat{\boldsymbol{\beta}}_{n,n}, \hat{\boldsymbol{\sigma}^2}_{n-1}, \hat{\boldsymbol{\gamma}}_n)$$

$$- \boldsymbol{V}_{2,n}(\boldsymbol{\beta}_n, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}) \ \big) \ \|\boldsymbol{\lambda}_{max}\|\big(\boldsymbol{M}_n^{-1}\big)$$

$$= \sum_{i=1}^{n}\Big(\ \frac{1}{E_i}exp\,(-\boldsymbol{X}_i\boldsymbol{\beta} + \frac{\boldsymbol{\sigma}^2}{2}\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2)\Big)$$

$$\leq \ \ n\ \mathbb{O}_p\Big(\ \sup_i|\boldsymbol{X}_i| \ \ \|\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n\|_2$$

$$+ \ \ |\hat{\boldsymbol{\sigma}^2}_n - \boldsymbol{\sigma}^2| \ \sup_i\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii} + \boldsymbol{\sigma}^2\sup_i|\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii} \ - \hat{\boldsymbol{V}}_{1,n}(\hat{\boldsymbol{\gamma}}_n)_{ii}| \ \ \Big)$$

$$= n \ \ \mathbb{O}_p\Big(\ \sup_i|\boldsymbol{X}_i| \ \ \|\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n\|_2 \ \Big)$$

$$= \mathbb{O}_p \ \ n \ \ \sqrt{(\frac{p_n}{n})}$$

Observing that order of last term dominates hence $\| \ \hat{\boldsymbol{\sigma}}_n^2 - \boldsymbol{\sigma}^2 \ \|_2 = \ \mathbb{O}_p \sqrt{(\frac{p_n}{n})}$ contrary to $\mathbb{O}_p \ (\frac{p_n}{n})$ due to contribution from last term. ■

the marginal method do not have the last term contributed through bias correction. Therfore the convergence rate for $\boldsymbol{\sigma}_n^2$ follows classical rate of $\mathbb{O}_p \ (\frac{p_n}{n})$

**Convergence of $\boldsymbol{\gamma}_n$**

$$\begin{aligned}\boldsymbol{\gamma} &= \frac{(\boldsymbol{Y}^* - \mathbb{X}\boldsymbol{\beta}_n)^\intercal \boldsymbol{M}_n \boldsymbol{W}_n (\boldsymbol{Y}^* - \mathbb{X}\boldsymbol{\beta}_n)}{(\boldsymbol{Y}^* - \mathbb{X}\boldsymbol{\beta}_n)^\intercal \boldsymbol{M}_n \boldsymbol{W}_n \boldsymbol{M}_n \boldsymbol{W}_n (\boldsymbol{Y}^* - \mathbb{X}\boldsymbol{\beta}_n) + (\boldsymbol{Y}^* - Y^{**})^\intercal \boldsymbol{M}_n \boldsymbol{W}_n \boldsymbol{M}_n \boldsymbol{W}_n (\boldsymbol{Y}^* - Y^{**})} \\ &= \frac{f_1(\boldsymbol{\beta}_n)}{f_2(\boldsymbol{\beta}_n) + C_n}\end{aligned}$$

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_n &= \frac{(\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n})^\intercal \boldsymbol{M}_n \boldsymbol{W}_n (\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n})}{(\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n})^\intercal \boldsymbol{M}_n \boldsymbol{W}_n \boldsymbol{M}_n \boldsymbol{W}_n (\boldsymbol{Y}^* - \mathbb{X}\hat{\boldsymbol{\beta}}_{n,n}) + (\boldsymbol{Y}^* - Y^{**})^\intercal \boldsymbol{M}_n \boldsymbol{W}_n \boldsymbol{M}_n \boldsymbol{W}_n (\boldsymbol{Y}^* - Y^{**})} \\ &= \frac{f_1(\hat{\boldsymbol{\beta}}_{n,n})}{f_2(\hat{\boldsymbol{\beta}}_{n,n}) + C_n} \\ &= \frac{f_1(\boldsymbol{\beta}_n) + 2(\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n)^\intercal \mathbb{X}^\intercal \boldsymbol{A}_n (\boldsymbol{Y}^* - \mathbb{X}\boldsymbol{\beta}_n) + (\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n)^\intercal \mathbb{X}^\intercal \boldsymbol{A}_n \mathbb{X} (\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n)}{f_2(\boldsymbol{\beta}_n) + 2(\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n)^\intercal \mathbb{X}^\intercal \boldsymbol{A}_n \boldsymbol{A}_n^\intercal (\boldsymbol{Y}^* - \mathbb{X}\boldsymbol{\beta}_n) + (\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n)^\intercal \mathbb{X}^\intercal \boldsymbol{A}_n \boldsymbol{A}_n^\intercal \mathbb{X} (\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}_n) + C_n} \\ &= \frac{f_1(\boldsymbol{\beta}_n) + g_1(\boldsymbol{\beta}_n, \hat{\boldsymbol{\beta}}_{n,n})}{f_2(\boldsymbol{\beta}_n) + g_2(\boldsymbol{\beta}_n, \hat{\boldsymbol{\beta}}_{n,n}) + C_n}\end{aligned}$$

Therefore $(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma})$ can be written as

$$\begin{aligned}\frac{f_1(\boldsymbol{\beta}_n) + g_1(\boldsymbol{\beta}_n, \hat{\boldsymbol{\beta}}_{n,n})}{f_2(\boldsymbol{\beta}_n) + g_2(\boldsymbol{\beta}_n, \hat{\boldsymbol{\beta}}_{n,n}) + C_n} - \frac{f_1(\boldsymbol{\beta}_n)}{f_2(\boldsymbol{\beta}_n) + C_n} &= \frac{f_2 g_1 + C_n g_1 - f_1 g_2}{(f_2 + C_n)(f_2 + g_2 + C_n)} \\ &= \frac{g_1}{f_2 + g_2 + C_n} - \frac{g_2}{f_2 + g_2 + C_n} \ \ \frac{f_1}{f_2 + C_n}\end{aligned}$$

It is proven that

$$g_1(\boldsymbol{\beta}_n, \hat{\boldsymbol{\beta}}_{n,n}) \ \ \leq \ \ \mathbb{O}_p(p_n)$$

$$g_2(\boldsymbol{\beta}_n, \hat{\boldsymbol{\beta}}_{n,n}) \;\geq\; \mathbb{O}_p(p_n)$$

$$f_1(\boldsymbol{\beta}_n) = \mathbb{O}_p(n)$$

$$f_2(\boldsymbol{\beta}_n) = \mathbb{O}_p(n)$$

$$C_n = \mathbb{O}_p(n)$$

$\|(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma})\|_2 = \frac{\mathbb{O}_p(p_n)}{\mathbb{O}_p(n)} \;\; = \mathbb{O}_p(\frac{p_n}{n})$ By Slutsky's theoreom

**Convergence of precision matrix**

**Lemma 1.12.13.** Let $\boldsymbol{A}$ and $\boldsymbol{A}_n$ be sequences of invertible matrices of same order belong to Banach space over $\left(\mathbb{R}^{n\times n}, \|\ \|_2\right)$ such that
$\|\boldsymbol{A}_n - \boldsymbol{A}\|_2 = \mathbb{O}_p(r_n)$ then $\|\boldsymbol{A}_n^{-1} - \boldsymbol{A}^{-1}\|_2 = \mathbb{O}_p(r_n)$ for large $n$

*Proof.* $\|\boldsymbol{A}_n^{-1} - \boldsymbol{A}^{-1}\|_2 = \|\boldsymbol{A}_n^{-1}\ (\boldsymbol{A}_n - \boldsymbol{A})\ \boldsymbol{A}^{-1}\|_2 \quad \|\boldsymbol{A}\|_2 = \sqrt{\sum_i \boldsymbol{\lambda}_i^2}$ where $\boldsymbol{\lambda}_i$ is $i$th singular value of $\boldsymbol{A}$
$\|\boldsymbol{A}^{-1}\|_2 = \sqrt{\sum_i \frac{1}{\boldsymbol{\lambda}_i^2}}$ $\quad \boldsymbol{\lambda}_i \neq 0$ for any $i$ due to non singularity of $\boldsymbol{A}$ ■

Using above lemma on $\boldsymbol{A} = \boldsymbol{\sigma}^2 \quad \boldsymbol{A}_n = \hat{\boldsymbol{\sigma}}_n^2$
The Frobenius norm inequality $\|\boldsymbol{A}\|_2 \;\leq\; \sqrt{n}\ \|\boldsymbol{A}\|_*$ here $\boldsymbol{A}$ is $n \times n$ matrix and $\|\boldsymbol{A}\|_*$ denotes spectral norm of $\boldsymbol{A}$

$$\begin{aligned}
\|\hat{\boldsymbol{\sigma}}_n^{-2}\hat{\boldsymbol{V}}_{1,n}^{-1} - \boldsymbol{\sigma}^{-2}\boldsymbol{V}_{1,n}^{-1}\|_2 \leq\;& \hat{\boldsymbol{\sigma}}_n^{-2}\ \ ||(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma})||_2\ \ \|\boldsymbol{W}_n\|_2 + \ \ \|\hat{\boldsymbol{\sigma}}_n^{-2} - \boldsymbol{\sigma}^{-2}\|_2\ \ \|(\boldsymbol{M}_n - \boldsymbol{\gamma}\boldsymbol{W}_n)\|_2 \\
\leq\;& \mathbb{O}_p(\frac{p_n}{n})\ \ \mathbb{O}(\sqrt{n}) \quad + \quad \mathbb{O}_p\sqrt{(\frac{p_n}{n})}\ \ \mathbb{O}(\sqrt{n})\ \ (1-\boldsymbol{\gamma})\max_i \boldsymbol{w}_{i+} \\
\leq\;& \mathbb{O}_p\left(\sqrt{p_n}\right)
\end{aligned}$$

Hence $\|\hat{\boldsymbol{V}}_{1,n}-\boldsymbol{V}_{1,n}\|_2 \;=\; \mathbb{O}_p\left(\sqrt{p_n}\right)$ implying that $\|\;(\hat{\boldsymbol{V}}_{1,n})_{i,i}\;-(\boldsymbol{V}_{1,n})_{i,i}\;\|_2 \;=\; \mathbb{O}_p\left(\sqrt{\frac{p_n}{n}}\right)$ provided all diagonal elements $\hat{\boldsymbol{V}}_{1,n}$ and $\boldsymbol{V}_{1,n}$ are of same order.

Now all diagonal elements of $\hat{\boldsymbol{V}}_{1,n}^{-1}$ and $\boldsymbol{V}_{1,n}^{-1}$ are of same order as well as off- diagonal elements due to assumption 8 on $w_{i,j}$s . This property of same order for inverse matrices can be verified using law

$(\boldsymbol{V}^{-1})_{i,i} = \frac{\det(\boldsymbol{V}_{\{i,i\}^c})}{\det(\boldsymbol{V})}$

$\boldsymbol{V}_{\{i,i\}^c}$ represents sub matrix of $\boldsymbol{A}$ obtained by deleting $i^{th}$ row and $i^{th}$ colomn from $\boldsymbol{A}$

$$\boldsymbol{V}_{2,n} = Diag\left[\;\frac{1}{E_i}exp\left(-\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{\sigma}^2\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2\right)\right]$$

By Mean Value Theorem (multivariate)

$$\begin{aligned}
&\|\;\hat{\boldsymbol{V}}_{2,n}\;-\;\boldsymbol{V}_{2,n}\|_2\\
&\leq n\;\max_i\frac{1}{E_i}exp\left(-\boldsymbol{X}_i\boldsymbol{\beta}_*+\boldsymbol{\sigma}^2\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2\right)\max_i\|\boldsymbol{X}_i\|_2\|(\hat{\boldsymbol{\beta}}_{n,n}-\boldsymbol{\beta}_n)\|_2\\
&+n\;\;\max_i\frac{1}{E_i}exp\left(-\boldsymbol{X}_i\boldsymbol{\beta}_*+\boldsymbol{\sigma}_*^2\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2\right)\quad\max_i(\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2)\;\;|\hat{\boldsymbol{\sigma}^2}_n-\boldsymbol{\sigma}^2|_2\\
&+n\;\;\max_i\frac{1}{E_i}exp\left(-\boldsymbol{X}_i\boldsymbol{\beta}_*+\boldsymbol{\sigma}_*^2\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2\right)\;\;\|\boldsymbol{\sigma}_*^2\;\;F(\boldsymbol{M}_n,\boldsymbol{W}_n,\boldsymbol{\gamma}_*)_{\;ii}/2\|_2\;\;\|(\hat{\boldsymbol{\gamma}}_n-\boldsymbol{\gamma})\|_2\\
&\leq \mathbb{O}_p\left(\sqrt{p_n}\right)
\end{aligned}$$

$$\begin{aligned}
\|\hat{\boldsymbol{V}^{-1}_{3,n}}\;-\boldsymbol{V}_{3,n}\|_2 \leq\;&\|\;\hat{\boldsymbol{\sigma}}_n^2-\boldsymbol{\sigma}^2\;\|_2\;\|\boldsymbol{V}_{2,n}\|_2\;+\;\|\hat{\boldsymbol{\sigma}}_n^2\|_2\;\;\|\hat{\boldsymbol{V}}_{2,n}\;-\;\boldsymbol{V}_{2,n}\|_2\\
&+\;\|\;\hat{\boldsymbol{V}}_{1,n}\;-\;\boldsymbol{V}_{1,n}\|_2
\end{aligned}$$

Using lemma 1.12.13 finally it is established that $\|\hat{\boldsymbol{V}^{-1}_{3,n}}-\boldsymbol{V}_{3,n}^{-1}\|_2 \;=\; \mathbb{O}_p\left(\sqrt{p_n}\right)$

**derivative wrt $\boldsymbol{\gamma}$**

$$\begin{aligned}
&\frac{\partial \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \\
&\qquad = -\boldsymbol{V}_{1,n}^{-1} \ \frac{\partial \boldsymbol{V}_{1,n}^{-1}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \ \boldsymbol{V}_{1,n}^{-1} \\
&\qquad = \boldsymbol{V}_{1,n}^{-1} \ \boldsymbol{W}_n \ \boldsymbol{V}_{1,n}^{-1} \\
&\qquad = \boldsymbol{M}_n^{-1}\boldsymbol{W}_n\boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\ \boldsymbol{M}_n^{-1}\boldsymbol{W}_n^2 - \boldsymbol{\gamma}\ \boldsymbol{W}_n^2\boldsymbol{M}_n^{-1} + \boldsymbol{\gamma}^2\ \boldsymbol{W}_n^3
\end{aligned}$$

Since gradient is a linear operator

$$\frac{\partial \big(\boldsymbol{V}_{1,n}\big)_{i,i}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \left(\frac{\partial \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\right)_{i,i}$$

$$(\boldsymbol{V}_{3,n}) = (\ \boldsymbol{\sigma}^2\ \boldsymbol{V}_{1,n} + \ \boldsymbol{V}_{2,n})$$

**Joint Convexity of Quasilikelihood** To estabilish convexity of quasilikelihood, we need to show the directional derivative for every direction

$$t \in \mathbf{R}^2 \quad < t, \nabla_{(\boldsymbol{\gamma}\ ,\boldsymbol{\sigma}^2)} S_n \geq 0$$

It can be established that the derivative of score function is monotonic

$$\begin{aligned}
\frac{\partial S_n}{\partial \boldsymbol{\gamma}} &= \mathbb{X}^\intercal \frac{\partial \big[\boldsymbol{V}_{\mathbf{3,n}}^{-\mathbf{1}}(\boldsymbol{\gamma})\big]}{\partial \boldsymbol{\gamma}} (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}) \\
&= \mathbb{X}^\intercal \frac{\partial \big[\ \boldsymbol{\sigma}^2\ \boldsymbol{V}_{1,n} + \ \boldsymbol{V}_{2,n}\big]^{-1}}{\partial \boldsymbol{\gamma}} (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})
\end{aligned}$$

**Derivative computation**

$$
\begin{aligned}
&\frac{\partial \boldsymbol{V}_{\mathbf{3,n}}^{-\mathbf{1}}}{\partial \boldsymbol{\gamma}} \\
&= -\boldsymbol{V}_{3,n}^{-1} \frac{\partial \boldsymbol{V}_{\mathbf{3,n}}}{\partial \boldsymbol{\gamma}} \boldsymbol{V}_{3,n}^{-1} \\
&= -\boldsymbol{\sigma}^2\, \boldsymbol{V}_{3,n}^{-1} \frac{\partial \boldsymbol{V}_{\mathbf{1,n}}}{\partial \boldsymbol{\gamma}} \boldsymbol{V}_{3,n}^{-1} \;-\; \boldsymbol{V}_{3,n}^{-1} \frac{\partial \boldsymbol{V}_{\mathbf{2,n}}}{\partial \boldsymbol{\gamma}} \boldsymbol{V}_{3,n}^{-1} \\
&= -\boldsymbol{\sigma}^2\, \boldsymbol{V}_{3,n}^{-1} \frac{\partial \boldsymbol{V}_{\mathbf{1,n}}}{\partial \boldsymbol{\gamma}} \boldsymbol{V}_{3,n}^{-1} \\
&- \boldsymbol{\sigma}^2\, \boldsymbol{V}_{3,n}^{-1}\, \mathrm{Diag}\Big( \frac{\big(\boldsymbol{V}_{2,n}\big)_{i,i}}{2} \frac{\partial \big(\boldsymbol{V}_{1,n}\big)_{i,i}}{\partial \boldsymbol{\gamma}} \Big) \boldsymbol{V}_{3,n}^{-1} \\
&= -\boldsymbol{\sigma}^2\, \boldsymbol{V}_{3,n}^{-1} \\
&\Big[ \mathrm{Diag}\Big( \exp\Big( \frac{\boldsymbol{\sigma}^2\, \boldsymbol{V}_{1,n\; i,i}}{2} \big(\boldsymbol{V}_{1,n}^{-1} W \boldsymbol{V}_{1,n}^{-1}\big)_{i,i} \Big) + \boldsymbol{V}_{1,n}^{-1} W \boldsymbol{V}_{1,n}^{-1} \Big] \boldsymbol{V}_{3,n}^{-1} \\
&\leq -\boldsymbol{\sigma}^2\, \boldsymbol{V}_{3,n}^{-1} \\
&\Big[ \mathrm{Diag}\Big( \exp\Big( \frac{\boldsymbol{\sigma}^2}{2(\boldsymbol{V}_{1,n}^{-1})_{i,i}} \big(\boldsymbol{V}_{1,n}^{-1} W \boldsymbol{V}_{1,n}^{-1}\big)_{i,i} \Big) + \boldsymbol{V}_{1,n}^{-1} W \boldsymbol{V}_{1,n}^{-1} \Big] \boldsymbol{V}_{3,n}^{-1} \\
&\leq -\boldsymbol{\sigma}^2\; \boldsymbol{V}_{3,n}^{-1} L(W, \boldsymbol{\gamma}) \boldsymbol{V}_{3,n}^{-1}
\end{aligned}
$$

**Remark 3.** The Quasi Likelihood $\int_{\beta}^{Y} S_n(\beta) d\beta$ depends on convexity of matrix $L(W, \boldsymbol{\gamma})$. However function of $\boldsymbol{\gamma}$ since adjacency matrix $W$ is not positive semi definite because by Perron Frobeius theorem it has at least one positive eigenvalue and some eigenvalues are negative since $\mathbf{trace}(W_n) = 0$

**Joint convexity**

**Lemma 1.12.14.** The unpenalized quasi likelihood function to be jointly convex function of $\boldsymbol{X\beta}\;\; \boldsymbol{\sigma}_2\;\; \boldsymbol{\gamma}$. The sufficient condition that score function is monotonic function of $\boldsymbol{X\beta},\;\; \boldsymbol{\sigma}_2,\;\; \boldsymbol{\gamma}$

*Proof.* Results that will be used in this proof are:

- If $V$ is non negative definite matrix then the function $f : V \to V$ is monotone

- The composition of two convex function is convex i.e $f, g$ are both monotone implies $f(g)$ is convex

- If $f, g$ are both monotone and f(x) is semi definite for all x $f \succeq 0$ and $g \succeq 0$ the product$fg$ is monotone.

$\boldsymbol{V}_{1,n}^{-1}(\boldsymbol{\gamma}) = (\boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\boldsymbol{W}_n)$ is monotonic function of $\boldsymbol{\gamma}$. So $\boldsymbol{V}_{1,n}$ is monotonic function of $\boldsymbol{\gamma}$ by result 1 $\boldsymbol{\sigma}^2\boldsymbol{V}_{1,n}$ is jointly convex function of and $\boldsymbol{\sigma}$ result 3. $\boldsymbol{V}_{2,n} = Diag\left( \frac{1}{E_i} exp\left(-\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\sigma}^2\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2\right)\right)$ is jointly convex function of $\boldsymbol{X}\boldsymbol{\beta}$ ,$\boldsymbol{\gamma}$ and $\boldsymbol{\sigma}$ by result 2 and 3. $\boldsymbol{V}_{1,n} + \boldsymbol{V}_{2,n}$ is jointly convex function of $\boldsymbol{X}\boldsymbol{\beta}$ ,$\boldsymbol{\gamma}$ and$\boldsymbol{\sigma}$. $\boldsymbol{V}_{3,n}^{-1} = (\boldsymbol{V}_{1,n} + \ \boldsymbol{\sigma}^2 \ \ \boldsymbol{V}_{2,n})^{-1}$ is jointly convex function of $\boldsymbol{X}\boldsymbol{\beta}$ ,$\boldsymbol{\gamma}$ and $\boldsymbol{\sigma}$ . $\underbrace{\frac{1}{n}\mathbb{X}^\intercal\left[\boldsymbol{\sigma}^2\mathbf{V_{3,n}}(\boldsymbol{\gamma})\right]^{-1}(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})}_{S_n}$ is jointly convex function of $\boldsymbol{X}\boldsymbol{\beta}$ ,$\boldsymbol{\gamma}$ and $\boldsymbol{\sigma}$ ■

**Central Limit Theorem**

1 We have link function such that

$$\mathbb{E}(\boldsymbol{Y}_i) = exp\left(\boldsymbol{X}_i^\intercal\boldsymbol{\beta} + \boldsymbol{\sigma}^2\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2\right) = \mu_i = h^{-1}(\boldsymbol{X}_i^\intercal\boldsymbol{\beta})$$

The $\frac{\partial h^{-1}}{\partial \boldsymbol{\beta}} < \infty$ and $\frac{\partial^2 h^{-1}}{\partial \boldsymbol{\beta}^2} < \infty$ for all $\boldsymbol{\beta}$ in parameter space

These conditions are satisfied since we assume $\|\boldsymbol{X}_i^\intercal\boldsymbol{\beta}\|, \infty$ , $\max_{i,j}(\mathbb{X}_{i,j})$ is bounded **??** and $h^{-1}(\boldsymbol{X}_i^\intercal\boldsymbol{\beta}) = exp(\boldsymbol{\sigma}^2\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}/2)exp(\boldsymbol{X}_i^\intercal\boldsymbol{\beta})$

2 there exists smooth variance function $V$ $VAR(Y_i|\mathbb{X}) = V(h^{-1}(\boldsymbol{X}_i^\intercal\boldsymbol{\beta}))$ This condition is satisfied since here $V_i(r) = r + r^2(e^{\boldsymbol{\sigma}^2\boldsymbol{V}_{1,n}(\boldsymbol{\gamma})_{ii}} - 1)$

3 Each element of vector of $\frac{\partial \boldsymbol{\mu}}{\partial \tilde{\boldsymbol{\beta}}}^\intercal\left[\mathbf{V_n}(\tilde{\boldsymbol{\sigma}}^{\mathbf{2}}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}_{\mathbf{n}})\right]^{-1}$ is finite

i.e $\|\mathbb{X}^\intercal Diag(\boldsymbol{\mu}_i)VAR^{-1}(\boldsymbol{Y})\|_\infty = \mathbb{O}(1)$ this is proved by following lemma.

**Lemma 1.12.15.** $\|\mathbb{X}^\intercal Diag(\boldsymbol{\mu}_i)VAR^{-1}(\boldsymbol{Y})\|_\infty < \infty$

*Proof.* Using assumptions which gives bounds on elements of $\mathbb{X}$ and parameters. Write D as diagonal matrix with $D_i = Var(Y_i)$

$$V_n(\boldsymbol{\sigma}^2, \boldsymbol{\gamma}, \boldsymbol{\beta}_n) = D' \, CORR(\boldsymbol{Y}) \, D$$

using Woodbury matrix identity $(A+A'BA)^{-1} = A^{-1} - (B^{-1}+A)^{-1}$ and taking $A = Diag(\boldsymbol{\mu}_i)$ with diagonal entries $0 > \boldsymbol{\mu}_i > \infty$ it is sufficient to prove that max of row sum of inverse of correlation matrix $\|CORR(\boldsymbol{Y})^{-1}\|_{1,\infty} < C$ bounded by some constant using lemma below $\|CORR(\boldsymbol{Y})\|_{1,\infty} < 1$ ■

**Peligrad's result on CLT** Define $(Y_i') = \Big(Y_i - \mu_i(\boldsymbol{\beta})\Big)$. $\mathcal{F}_1^m$ = Sigma field$(Y_i' : 1 \leq i \leq m)$ Let $T_n = \sum_{i=1}^n Y_i'$ and $\nu_n^2 = \mathbb{E}(T_n)^2$

$$\rho(n) = \sup_{\substack{m \geq 0 \\ X \in \mathbf{L}_2(\mathcal{F}_1^m) \\ Y \in \mathbf{L}_2(\mathcal{F}_{m+n}^\infty)}} |CORR(X,Y)| \quad \alpha(n) = \sup_{\substack{m \geq 0 \\ A \in \mathcal{F}_1^m \\ B \in \mathcal{F}_{m+n}^\infty}} |P(A \cap B) - P(A)P(B)|$$

Peligard showed that for non stationary sequences $\{Y_i'\}_{i=1}^\infty$ CLT is obtainable provided some sufficient conditions and and $\lim_{n\to\infty} \rho(n) < 1$ and $\lim_{n\to\infty} \alpha(n) = 0$ Using the methods we can show CLT holds under these weak sufficient condition.

1 $\mathbb{E}(Y_i) = 0$ $\mathbb{E}(Y_i^2) < \infty$

2 $\max_{1 \leq i \leq n} \mathbb{E}|Y_i'|^{2+\delta} < \infty$ Lyapunov condition which suffices for Linderberg Conditions

3 $\nu_n^2 = \mathbf{O}\Big(nh(n)\Big)$ $h(n)$ is slow varying function i.e $\lim_{n\to\infty} \frac{h(an)}{n} = 1$ for all $a > 0$

1.1 In our setup $\sup_i \boldsymbol{V}_{1,n\,i,i+n}^{-1} = 0$ implies under Gaussian setup $\alpha(n) \to 0$ thus making $\Psi_i$ and $\Psi_{i+n}$ ae independent hence $Y_i$ and $Y_{i+n}$ are independent as as n tends to infinity. hence $\alpha(n) \to 0$ when under probability space corresponds to process $Y_i$

1.2 $Corr(Y_i, Y_j) \leq Corr(\Psi_i, \Psi_j) = \frac{V_{1,n\,i,j}}{\sqrt{V_{1,n\,i,i}\, V_{1,n\,j,j}}} \leq |\boldsymbol{\gamma}|$ for all $i, j$ so $\lim_{n\to\infty} \rho(n) < 1$

2 $\delta = 1$ is satisfied here

3 $\sup_i(\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^2 \boldsymbol{\sigma}^2 \frac{1}{w_{i+}(1+|\boldsymbol{\gamma}|)}\ ) \leq h(n) \leq \sup_i \frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^2 \left( exp(\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n\ \ i,i}) - 1 \right)}{\boldsymbol{\sigma}^2 \boldsymbol{V}_{1,n\ \ i,i}}\ \frac{1}{w_{i+}(1-|\boldsymbol{\gamma}|)}$

$h(n)$ is independent of $n$ thus regular varying function of $n$.

**Implication of CLT** The unpenalized score equation $S_n(\boldsymbol{\beta}_n, \boldsymbol{\sigma}_2, \boldsymbol{\gamma})$ converges to Normal distribution when parameters assume values of true model. Thus if can show that solution of score equation i.e $S_n(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\sigma}_2, \boldsymbol{\gamma}) = \mathbf{0}$ is consistent to $\boldsymbol{\beta}_{n,0}$ then $(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0})$ converges weakly to Normal distribution. Further when conditions are satisfied Liang and Zeger the solutions of $S_n(\hat{\boldsymbol{\beta}}_n^*, \hat{\boldsymbol{\sigma}}_2, \hat{\boldsymbol{\gamma}}) = \mathbf{0}$Liang & Zeger (1986) follows CLT i.e $(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) \longrightarrow \mathbf{N}$

**Exact structure Of V** If the process is stationary the exact form of $V^{-1}$ is known through paper Toeplitz and Circulant Matrices: A review . When not stationary can we know analytical structure? For example let $V_n(\boldsymbol{\gamma}) = (M_n^{-1} - \boldsymbol{\gamma}\, W_n)^{-1}$ where number of '1's in row of matrix $W_n$ is less than $k$ We can perform update $V_n(\boldsymbol{\gamma}_{t+1}) = V_n(\boldsymbol{\gamma}_t) + (\boldsymbol{\gamma}_{t+1} - \boldsymbol{\gamma}_t) \frac{\partial \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} | \boldsymbol{\gamma}_t$

$$
\begin{aligned}
&\frac{\partial \boldsymbol{V}_{1,n}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \\
&= -\boldsymbol{V}_{1,n}^{-1}\ \frac{\partial \boldsymbol{V}_{1,n}^{-1}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\ \boldsymbol{V}_{1,n}^{-1} \\
&= \boldsymbol{V}_{1,n}^{-1}\ \boldsymbol{W}_n\ \boldsymbol{V}_{1,n}^{-1} \\
&= \boldsymbol{M}_n^{-1} \boldsymbol{W}_n \boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\ \boldsymbol{M}_n^{-1} \boldsymbol{W}_n^2 - \boldsymbol{\gamma}\ \boldsymbol{W}_n^2 \boldsymbol{M}_n^{-1} + \boldsymbol{\gamma}^2\ \boldsymbol{W}_n^3
\end{aligned}
$$

Therefore diagonal elements are $diag(\boldsymbol{M}_n^{-1} \boldsymbol{W}_n \boldsymbol{M}_n^{-1} - \boldsymbol{\gamma}\ \boldsymbol{M}_n^{-1} \boldsymbol{W}_n^2 - \boldsymbol{\gamma}\ \boldsymbol{W}_n^2 \boldsymbol{M}_n^{-1} + \boldsymbol{\gamma}^2\ \boldsymbol{W}_n^3)_i = 0 - 2\boldsymbol{\gamma} w_{+i}^2 + \boldsymbol{\gamma}^2 W_{i,i}^3 \leq -2\boldsymbol{\gamma} k_{+i}^2 + \boldsymbol{\gamma}^2 k(k-1)/2$

# Chapter 2

# High Dimensional Sparse-$LDA$ for spatio-temporal Data

## Introduction

## 2.1 Introduction

The mathematical clarity and simplicity of Fisher's linear discriminant analysis (LDA) has lead to extensive applications in multimedia information retrieval such as speech and pattern recognition as explained by Yu & Yang (2001). Used both for classification Cox & Savoy (2003) Gutman et al. (2013) and dimension reduction Mourao-Miranda et al. (2005) Rathi & Palani (2012) , in the context of image analysis and feature extraction, LDA has often encountered difficulty pertaining to such high-dimensional problems. Technolgical advancements in medical images specifically magnetic resonance images (MRIs) have now lead to a rise in high-resolution images with dimensions as high as 256x256x198 $\approx$ 12M volumetric pixels or voxels. As the number of subjects obtained to study such images cannot feasibly exceed these dimensions, it leads to singularity in the construction of the covariate matrix. An initial proposition of a simple two-step algorithm such as PCA-LDA used yet another dimension reductionality step by reducing the initial dimension using principal component analysis (PCA) Wang et al. (2010) Ghosh (2001). Although this technique leads to construction

of orthogonal features of lower dimensions, the components from LDA and PCA algorithms maybe incompatible with each other, thus resulting in the loss of important information.

The review provided by (Fan & Lv, 2010, Section 4.2) provides insight into issues that classical LDA encounters when a high-dimensional problem arises. They show that dimension reduction significantly is important for reducing the misclassification rate. Certain propositions considered remedies such as imposing independence assumptions on the covariance structure thus significantly lowering the number of estimating parameters to circumvent the singularity Bickel et al. (2004),Tibshirani et al. (2002) and Fan & Fan (2008). As an application to genetics these techniques did not necessitate the selection of features to facilitate the classification. Fan & Fan (2008) produced the method named Features Annealed Independence Rule (FAIR). This was an improvement over nearest shrunken centroid rule Tibshirani et al. (2003) essentially equivalent to two-sample t-tests, as it sets a relative importance order for features that would result in a more optimal selection. This procedure keeps in check the noise accumulation previously unaccounted for, so as to not subvert faint features. However in an attempt to incorporate and account for significant correlation among the genes, Fan et al. (2012) proposed the regularized optimal affine discriminant (ROAD) and a few variations in nthe assumptions made. For the same microarray dataset J. Shao et al. (2011) propose a sparse LDA (SLDA) using thresholding obtain a sparse estimate of the covariance matrix. Another important contribution with regard to LDA classification in genetics was made by Witten & Tibshirani (2011). The authors here penalized the discriminant vectors in Fisher's discriminant problem. This results in constraining the within subject covariance while penalizing the between subject covariances. Instead of separately estimating or penalizing the covariance estimates or mean estimates, Cai & Liu (2011) propose sparse estimates of the product of the difference in class means and the covariance, an important portion of the classification rule using $l_1$ minimization. In other work, Mai et al. (2012) propose penalized least squares estimate using a lasso penalty to solve the high dimensional LDA problem.

For longitudinal neuroimaging studies data has naturally structured dependencies in higher dimensions that may significantly inform the classification rule. This article explores statistical methodology that may assist in the simultaneous selection of regions and classification of diseases status using structural brain magnetic resonance images (MRIs). The objective would be able to assist researchers in locating brain regions or ROI (Region of Interest) based analysis that identifies specific voxels playing a key role in investigating regions of susceptibility to Alzheimer's disease. The authors of Yingjie & Maiti (2019) explored penalized LDA imposing a parametric spatial covariance on the within subject images. We extend this work to investigate methods using a spatio-temporal covariance for longitudinal studies and explore estimation under a variety of situations such as space-time separability and non-separability in section 2.4 and covariance tapering in section 2.8.1. All of the methodology in this article has been explored under the smoothly clipped absolute deviation (SCAD) penalty Fan & Li (2001). The algorithm use to obtain the penalized parameters was introduced by Zou & Li (2008) as the one-step sparse estimates. The consistency and selection properties have been studied under this one-step setup. We produce a series of simulation results under a variety of setups in section A dataset from the Alzheimer's disease neuroimaging initiative (ADNI Petersen et al. (2010)) is used to demonstrate the method and results in areas of the hippocampus that may play a vital role in the classification of healthy brains and patients diagnosed with AD.

## 2.2 Review of classical Linear discriminant Analysis (LDA)

Consider a $pT$-dimensional discriminant problem between two classes $\mathcal{C}_1$ and $\mathcal{C}_2$. Let $Y_{k1}, .., Y_{kn_k}$ be from classes $\mathcal{C}_k$, where $k \in \{1, 2\}$ and $Y_{kj} \in \mathbb{R}^{pT}$, We further assume that $Y_{kj} \sim N_{pT}(\mu_k, \Sigma(\theta))$ are independent and identically distributed for $j = \{1, 2, .., n_k\}$. The mean vectors $\mu_k$ vary between

the classes but they have a common variance $\Sigma(\theta)$ with parameter $\theta$. The estimation is based on samples of replicates of size $n_1$ and $n_2$ for each class respectively and the total sample size is $n = n_1 + n_2$. The membership of a new test sample $X$ into class $\mathcal{C}_1$ is then determined by the LDA classifier given by $\hat{\delta}(X)$ such that,

$$\hat{\delta}(X) = \left(X - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)\right)\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2) > 0 \tag{2.1}$$

Alternatively this classifier can be expressed in terms of the difference in mean $\Delta = \mu_1 - \mu_2$ that provides insight into the discriminant vector.

$$\hat{\delta}(X) = \left(X - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)\hat{\Sigma}^{-1}\hat{\Delta}\right) > 0 \tag{2.2}$$

where $\hat{\mu_k}$ and $\hat{\Sigma}$ denote the estimated mean and covariance. The maximum likelihood estimates (MLE) that results in sample means and covariances can be obtained by maximizing the log-likelihood function for $\mu_k$ and $\theta$ is given by,

$$\begin{aligned}(\hat{\mu}_k, \hat{\theta}) = \arg\max_{\mu_k,\theta} \mathcal{L}(\theta, \mu_1, \mu_2; Y) = &-\frac{\sum_{t=1}^{T} pT \times n}{2} log(2\pi) - \frac{n}{2} log|\Sigma(\theta)| \\ &- \frac{1}{2}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(Y_{k,i} - \mu_k)^T\Sigma^{-1}(\theta)(Y_{k,i} - \mu_k)\end{aligned} \tag{2.3}$$

All the estimates obtained in the setup have consistent properties under the setup where $pT < n$. This Bayes' classifier rule is established within the parameter space where $l_1$ and $l_2$ are positive constants such that,

$$\Gamma = \{(\Delta, \Sigma(\theta)) : \Delta^T\Sigma^{-1}(\theta)\Delta > \mathbb{C}_{pT}, l_1 \leq \lambda_{min}(\Sigma(\theta)) \leq \lambda_{max}(\Sigma(\theta)) \leq l_2\} \tag{2.4}$$

If the new observation $X$ belongs to $\mathcal{C}_1$, then the conditional misclassification rate for unknown parameters $\Theta = (\mu_1, \mu_2, \theta)$ of $\hat{\delta}(X)$ is given by,

$$\mathbb{W}_1(\hat{\delta}, \Theta) = P(\hat{\delta}(X)) \leq 0 | X \in \mathcal{C}_1) = 1 - \Phi(\psi_1) \tag{2.5}$$

where,

$$\psi_1 = \frac{\frac{1}{2}(\hat{\mu_1} - \hat{\mu_2})^T \hat{\Sigma} \hat{\Delta}}{\sqrt{\hat{\Delta}^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} \hat{\Delta}}} \tag{2.6}$$

Similarly the expression for $X$ belonging to $\mathcal{C}_2$ results in $\mathbb{W}_2(\hat{\Delta}, \Theta) = 1 - \Phi(\psi_2)$. Therefore the overall misclassification rate is given by,

$$\mathbb{W}(\hat{\delta}) = \max_{\Theta \in \Gamma} \frac{1}{2} (\mathbb{W}_1(\hat{\Delta}, \Theta) + \mathbb{W}_2(\hat{\Delta}, \Theta)) \tag{2.7}$$

Under the assumption that $X$ is normally distributed, the worst case conditional classification error rate where $\Phi(\cdot)$ denotes the standard Gaussian distribution function is given by,

$$\mathbb{W}(\delta) = \max_{\Theta \in \Gamma} [1 - \Phi(\frac{\sqrt{\Delta^T \Sigma^{-1} \Delta}}{2})] = 1 - \Phi(\frac{\sqrt{\mathbb{C}_{pT}}}{2}) \tag{2.8}$$

In the neuroimaging setup with longitudinal structural MRIs, as discussed in section 2.1, we encounter a high dimensional sceniario where $pT >> n$ rendering singularity of covariance matrix estimates $\hat{\Sigma(\theta)}$. In what follows, we impose a spatio-temporal dependence between registered images of each subject.

## 2.3 Spatio-temporal LDA

For the purposes of classification, the voxels present in MRI images of a particular subject is modeled as Gaussian Random Field (GRFs). These random fields exhibit decaying spatial covariance dependences in MRIs for each particular time point for a subject. Hence, longitudinal MRIs of a specific subject can be considered to posses spatio-temporal dependence. To begin with, we consider a separable covariance structure where both spatial covariance on each image and the time covariance across images within the same subject are assumed to be second order stationary and isotropic and identical across subjects.

Let us consider a $p \cdot T$ dimensional discriminant problem of a spatio-temporal process between two classes $\mathcal{C}_1$ and $\mathcal{C}_2$. Let $\{Y_{ki}(s,t) : (s,t) \in D \times \mathcal{T}\}$, for $i = 1, .., n_k$ in each class $k = \{1, 2\}$ of size $n_1$ and $n_2$ respectively such that $n_1 + n_2 = n$, denote a spatio-temporal process in the domain $D \times \mathcal{T} \in \mathbb{R}^d \times [0, \infty]$ where $d$ denotes the spatial dimensionality. Also assume, $\frac{n_1}{n} \to \pi$ for $0 < \pi < 1$ an $n \to \infty$. Let us express the process $Y_{kj}$ as,

$$Y_{ki}(s,t) = \mu_k(s,t) + \epsilon_{ki}(s,t) \tag{2.9}$$

where $\mu_k$ is the mean effect corresponding to each class $k = \{1, 2\}$ and the error term $\{\epsilon_{ki}(s,t) : (s,t) \in D \times \mathcal{T}\}$ is a Gaussian process with mean 0 and covariance $\Sigma(\theta)_{(s,t),(s',t')} = \gamma[(s,t),(s',t');\theta]$ such that,

$$\gamma[(s,t),(s',t');\theta] = cov[\epsilon(s,t), \epsilon(s',t') = \gamma[\, \|(s-s')\|_2, |t-t'|;\theta] \tag{2.10}$$

We additionally constrain the the spatio-temporal voxels to be non-random, and for any pair of spatio-temporal sites the distance is bounded below by fixed number such that $\|(s,t) - (s',t')\| \geq \eta > 0$. Hence all of the statistical methodology described will be investigated under the increasing

domain framework.

Alternatively, the vectorized form of the model can be expressed as,

$$Y_{k,i,j,l} = \mu_{k,j,l} + \epsilon_{k,i,j,l} \tag{2.11}$$

where, $j = 1, 2, ...p$ represents the $s_j^{th}$ spatial site, $l = 1, 2, ...T$ represents the $t_l^{th}$ time point. Specifically, $(Y_{k,i})_{pT\times 1} = (Y_{k,i,1,1}, ...., Y_{k,i,p,T})'$ , $(\mu_k)_{pT\times 1} = (\mu_{k,1,1}, ...., \mu_{k,p,T})'$ and $\epsilon_{k,i} = (\epsilon_{k,i,1,1}, ...., \epsilon_{k,i,p,T})'$ has a multivariate Gaussian distribution. Thus, $Y_{k,i} \sim \mathbf{N}_{pT}(\mu_k, \Sigma(\theta))$ where $\Sigma(\theta)$ is $pT \times pT$ covariance matrix.

## 2.4 Spatio-temporal LDA

For the purposes of classification, the voxels present in MRI images of a particular subject is modeled as Gaussian Random Field (GRFs). These random fields exhibit decaying spatial covariance dependences in MRIs for each particular time point for a subject. Hence, longitudinal MRIs of a specific subject can be considered to posses spatio-temporal dependence. To begin with, we consider a separable covariance structure where both spatial covariance on each image and the time covariance across images within the same subject are assumed to be second order stationary and isotropic and identical across subjects.

Let us consider a $p \cdot T$ dimensional discriminant problem of a spatio-temporal process between two classes $\mathcal{C}_1$ and $\mathcal{C}_2$. Let $\{Y_{ki}(s,t) : (s,t) \in D \times \mathcal{T}\}$, for $i = 1, .., n_k$ in each class $k = \{1, 2\}$ of size $n_1$ and $n_2$ respectively such that $n_1 + n_2 = n$, denote a spatio-temporal process in the domain $D \times \mathcal{T} \in \mathbb{R}^d \times [0, \infty]$ where $d$ denotes the spatial dimensionality. Also assume, $\frac{n_1}{n} \to \pi$ for $0 < \pi < 1$ an $n \to \infty$. Let us express the process $Y_{kj}$ as,

$$Y_{ki}(s,t) = \mu_k(s,t) + \epsilon_{ki}(s,t) \tag{2.12}$$

where $\mu_k$ is the mean effect corresponding to each class $k = \{1, 2\}$ and the error term $\{\epsilon_{ki}(s,t) : (s,t) \in D \times \mathcal{T}\}$ is a Gaussian process with mean 0 and covariance $\Sigma(\theta)_{(s,t),(s',t')} = \gamma[(s,t),(s',t');\theta]$ such that,

$$\gamma[(s,t),(s',t');\theta] = cov[\epsilon(s,t),\epsilon(s',t') = \gamma[\,\|(s-s')\|_2, |t-t'|;\theta] \tag{2.13}$$

We additionally constrain the the spatio-temporal voxels to be non-random, and for any pair of spatio-temporal sites the distance is bounded below by fixed number such that $\|(s,t)-(s',t')\| \geq \eta > 0$. Hence all of the statistical methodology described will be investigated under the increasing domain framework.

Alternatively, the vectorized form of the model can be expressed as,

$$Y_{k,i,j,l} = \mu_{k,j,l} + \epsilon_{k,i,j,l} \tag{2.14}$$

where, $j = 1, 2, ...p$ represents the $s_j^{th}$ spatial site, $l = 1, 2, ...T$ represents the $t_l^{th}$ time point. Specifically, $(Y_{k,i})_{pT\times 1} = (Y_{k,i,1,1}, ...., Y_{k,i,p,T})'$ , $(\mu_k)_{pT\times 1} = (\mu_{k,1,1}, ...., \mu_{k,p,T})'$ and $\epsilon_{k,i} = (\epsilon_{k,i,1,1}, ...., \epsilon_{k,i,p,T})'$ has a multivariate Gaussian distribution. Thus, $Y_{k,i} \sim \mathbf{N}_{pT}(\mu_k, \Sigma(\theta))$ where $\Sigma(\theta)$ is $pT \times pT$ covariance matrix.

### 2.4.1 Spatio-temporal covariance

In this section, we begin by reviewing the property that any covariance matrix may retain a irreducible block structure. Let $\Sigma_{PT\times PT}$ be a covariance matrix which is formed by unfolding the covariance tensor $\Sigma$ in the following ordered way,

$$\Sigma_{P\times T\times P\times T}(a,a',c,c') = \Sigma_{PT\times PT}(Pa+c, Pa'+c')$$

Let $q = (P-1)a + c$ and $q' = (P-1)a' + c'$ then, $\Sigma(q, q') = Cov(Y(s_c, t_a), Y(s_{c'}, t_{a'}))$. Therefore a fully separable model would produce, $\Sigma(q, q')_{PT \times PT} = \gamma_1(s_c - s'_c; \theta)\gamma_2(t_a - t'_a; \theta)$ and a non-separable model would result in $\Sigma(q, q') = \gamma(\sqrt{(s_c - s'_c)^2 + (t_a - t'_a)^2}; \theta)$.

All covariance models used in this article is assumed to be up to second order stationary. Cressie & Huang (1999) describe in detail classes of spatio-temporal stationary covariance functions. Assume $\gamma(\cdot, \cdot)$ is continuous and its spectral distribution posses a spectral density $f(\omega, \tau) \geq 0$, that is by Bochner's theorem,

$$\gamma(s, t) = \int_{\mathbf{R}^d} \int_{\mathbf{R}} e^{i\omega' s + i\tau' t} f(\omega, \tau) d\omega d\tau \tag{2.15}$$

Additionally if $\gamma(\cdot, \cdot)$ is also integrable, we get

$$f(\omega, \tau) = \frac{1}{2\pi^{d+1}} \int \int e^{-is'\omega - it\tau} \gamma(s; t) ds dt = \frac{1}{2\pi} \int e^{-is\tau} h(\omega; t) dt \tag{2.16}$$

where $h(\omega; t) := \frac{1}{2\pi^d} \int e^{-is'\omega} \gamma(s; t) ds dt = \int e^{it\tau} f(\omega, \tau) d\tau$. A valid positive definite covariance can be achieved by modeling $h(\omega; u) = \rho(\omega; u)k(\omega)$ where the following is satisfied,

(C1) For each $\omega \in \mathbb{R}^d, \rho(\omega; \cdot)$ is a continuous autocorrelation function, $\int \rho(\omega; u) du < \infty$ and $k(\omega) > 0$.

(C2) $\int k(\omega) d\omega < \infty$.

Matérn (1960) defines a class of space time interacting model through the spectral density

$$f(\omega, \tau) = \eta(\alpha^2\beta^2 + \beta^2\|\omega\|^2 + \alpha^2\tau^2 + \epsilon\|\omega\|^2\tau^2)^{-(\nu + \frac{d+1}{2})} \tag{2.17}$$

for all $\eta, \beta, \alpha, \nu \geq 0$ and belonging to compact subspace and $\epsilon \in [0, 1]$ which represents the extent of separability between the time and space domain. $\alpha^{-1}, \beta^{-1}$ represents the spatial and temporal decay of the correlation respectively. $\eta$ is the scale parameter and $\nu$ is the smoothness parameter.

Additionally it can be shown that this Matern class of covariances satisfy the regularity conditions.

### 2.4.2 Irregular lattice points in space and time

Cressie & Lahiri (1996) gives general conditions for uniform convergence REML estimates for lattice as well irregular spaced points. Theses conditions are satisfied for large class of covariance structure including Matern class.However we assume that our brain is embedded in irregular spatial lattice and time points are at irregular lag.

Let us define the indexing set of time point $\{0, 1, 2, ..n\}$ and $n \to \infty$ . For each time point $t \in \mathcal{T}_n = \{0, t_1, ..t_n\}$ we have a spatial domain $\mathcal{S}_t$. Similarly we can define subset consisting of odd indices $\mathcal{T}_{1n} = \{0, t_1, ..t_{2k+1} : 0 \leq k \leq \lfloor n/2 \rfloor\}$ Here d=3 each $p_{i,t}$ for $i \in 1, 2, ..d$ representing the number of points in each direction.Here the cardinality $|S_t| = p_{1,t}p_{2,t}p_{3,t} = P_t$. We assume that the difference between spatial lattice is fixed which denoted by $H = (h_1, h_2, h_3)$ independent of $n$. Define the index set containing origin and neighbors at unit distance to be $\mathcal{L} = \{(l_1, l_2, ..l_d) : \sum_i^d |l_i| \leq 1\}$. Define $\mathbb{Z}$ by the set of all positive and negative integers including 0. $\mathbb{Z}_2 = \{2i; i \in \mathbb{Z}\}$ the set of all even integers. Let the generator of indexing set $Z_{n,t} = \{(i_1, ..i_j, ..i_d) : 0 \leq i_j \leq p_i \text{ and } 1 \leq j \leq d\}$ $Z_{1n,t} = \{(k_1, ..j_j, ..k_d) : k_j = 2x_j + 1\, 0 \leq x_j \leq 2\lfloor (p_i - 1)/2 \rfloor - 1 \text{ and } 1 \leq j \leq d\}$ be odd integer subset of $Z_n$ . Therefore regular spatial lattice is generated by $\mathcal{S}_t = \mathcal{L} \circ Z_{n,t} \circ H$ and $\mathcal{S}_{1t} = \mathcal{L} \circ Z_{1n,t} \circ H$

### 2.4.3 Well separateness in space and time domain

However stationarity,isotropy and non increasing as function of distance is satisfied by Matern covariance functions,this properties allows to validate our results when space does not have lattice structure. Henceforth, we assume our spatial points in irregular domain $\mathcal{S}_t$for all $t \in \{0, t_1, ..t_{T_n}\}$ only with the restriction $a \leq \inf_{s' \in \mathcal{S}_t} \|s' - s\|_2 \leq A$ for some constants $a, A > 0$ independent of time.

Similarly,for irregular time domain for all $\mathcal{T}_n$ have $b \leq \inf_{t \in \mathcal{T}_n} \|t' - t\| \leq B$ for some constants $b, B > 0$. Therefore we have an increasing domain setup.

## 2.5 Regularity Conditions

Denote $\sum_{t=1}^{T} P_t = \tilde{P}_T$ and $\sum_{t=1}^{T} s_t = \tilde{s}_T$ $\Sigma(\theta)$ is assumed to be second order stationary, isotropic and twice differential over all dimensions of space and time for $\theta \in \Xi$ over all $(s,t) \in D \times \mathcal{T}$ where $\Xi$ is the parametric space of $\theta$.

Define projection covariance matrix $\tilde{\Pi}(\theta) = \tilde{\Sigma}^{-1}(\theta) - \tilde{\Sigma}^{-1}(\theta)\tilde{X}_1(\tilde{X}_1^T\tilde{\Sigma}^{-1}(\theta)\tilde{X}_1)^-\tilde{X}_1^T\tilde{\Sigma}^{-1}(\theta)$

In general, let us consider any covariance matrix $\Sigma(\theta)$ for $\theta \in \Xi$ constructed by a covariance function $\gamma(\theta)$ where the true parameter $\theta$ is given by $\theta_0$. Also note that the derivative of $\Sigma(\theta)$ w.r.t $\theta_m$ is denoted by $\Sigma^m(\theta) = \frac{\partial}{\partial\theta_m}\Sigma(\theta)$ and second derivative is denoted by $\Sigma^{mm'}(\theta) = \frac{\partial^2\Sigma(\theta)}{\partial\theta_m\partial\theta_{m'}}$ where $\theta = (\theta_1, ..\theta_m, ..\theta_r)'$. Then we require the following assumptions,

(A1) $\frac{n_1}{n} \to \pi$

(A2) $\sum_{t=1}^{T} s_t = o(n)$

(A3) Let $\mathcal{V} =$ denote a $r \times r$ matrix so that $trace(\tilde{\Pi}(\theta)\frac{\partial\tilde{\Sigma}(\theta)}{\partial\theta_m}\tilde{\Pi}(\theta)\frac{\partial\tilde{\Sigma}(\theta)}{\partial\theta_{m'}}) = v_{m,m'}$ are the elements of $\mathcal{V}$, then $lim_{p\wedge T\to\infty}\frac{v_{m,m'}}{\sqrt{v_{mm}}\sqrt{v_{m'm'}}}$ exists and $\mathcal{V}$ is non-singular.

(A4) For any compact subset $K \subset \Theta$ have,any $\theta \in K$ and finte non zero constant $\zeta_{1,K}$ and $\zeta_{2,k}$ and for all $m, m' \in \{1, 2, ..r\}$ we have,

(A4).1 $\lim_{p\wedge T\to\infty} \lambda_{max}(\Sigma(\theta)) = O(1)$

(A4).2 $lim_{p\wedge T\to\infty}\lambda_{max}\Sigma^m(\theta_0) < \infty = O(1)$ where $\Sigma^m(\theta_0) = \frac{\partial\Sigma(\theta)}{\partial(\theta_m)}|_{\theta=\theta_0}$.

(A4).3 $lim_{p\wedge T\to\infty}\lambda_{max}\Sigma^{mm'}(\theta^*) < \infty = O(1)$ where $\Sigma^{mm'}(\theta^*) = \frac{\Sigma(\theta)}{\partial(\theta_m)\partial(\theta'_m)}|_{\theta=\theta^*}$.

(A4).4 $\|\Sigma^m(\theta^*)\|_F^2 = O((\sum_{t=1}^{T} P_t)^{\frac{1}{2}+\delta})$ where $\delta > 0$.

(A4).5 $\lim_{p\wedge T\to\infty} \lambda_{min}(\Sigma(\theta)) > \zeta_{2,K} > 0$

(A4).6 For some subsequence $u_n$ and $\delta > 0$ such that $\limsup_n \frac{u_n}{n} \geq 1 - \delta \liminf_{n\to\infty} \lambda_{r_n}\Sigma^m(\theta) > \zeta_{2,K} > 0$ for all $m$ where $\lambda_1 \geq \lambda_2 \geq ..\lambda_{\tilde{P}_T}$ .

### 2.5.1 Asymptotic optimal misclassification rate

Below under the usual setup for classical LDA, we show the behavior of the misclassification error rate. First lemma 2.5.4 provides the asymptotic matrix property below.

**Theorem 2.5.1.** Let $\hat{\theta}_n$ be Restricted Expected Maximum Likelihood(REML) estimate of $\theta_0$ then $\|\hat{\theta}_n - \theta_0\|_2 = O_p(\sqrt{\frac{1}{\sum_{t=1}^{T} P_t\, n}})$

*Proof.* Assuming conditions A(1)-A(4) hold, Cressie & Lahiri (1996) theorem 3.2 shows that

$$\mathcal{V}^{\frac{-1}{2}}(\hat{\theta}_{REML} - \theta_0) \Rightarrow N(\mathbf{0}, \mathbb{I})$$

Since $\mathcal{V}_{i,j} = trace(\tilde{\Pi}(\theta)\frac{\partial\tilde{\Sigma}(\theta)}{\partial\theta_m}\tilde{\Pi}(\theta)\frac{\partial\tilde{\Sigma}(\theta)}{\partial\theta_{m'}}) \geq O(\sum_{t=1}^{T} P_t\, n)\inf_i \lambda_i(\tilde{\Pi}(\theta)\frac{\partial\tilde{\Sigma}(\theta)}{\partial\theta_m}\tilde{\Pi}(\theta)\frac{\partial\tilde{\Sigma}(\theta)}{\partial\theta_{m'}})$ We infer that $\|\hat{\theta}_n - \theta_0\|_2 = O_p(\sqrt{\frac{1}{\sum_{t=1}^{T} P_t\, n}})$ Later we show that above assumptions holds for Materrn class covariance even in irregular space time domain ■

**Lemma 2.5.2.**

$$\Big(\frac{\hat{\Delta}^T\Sigma^{-1}\hat{\Delta}}{2\sqrt{\hat{\Delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\Delta}}} - \sqrt{\Delta^T\Sigma^{-1}\Delta}/2\Big) = \frac{O_P(\frac{PT}{n})(\sqrt{\Delta^T\Sigma^{-1}\Delta})}{\sqrt{\Delta^T\Sigma^{-1}\Delta)}(\sqrt{\Delta^T\Sigma^{-1}\Delta}) + O_P(\frac{PT}{n})}$$

*Proof.* From assumptions A(4), we get $\hat{\Sigma} = \Sigma + \epsilon E$ where $\epsilon = O_p(\sqrt{\frac{1}{PTn}})$

and $\|E\|_2 < \infty$

So by geometric series expansion is valid for large $n$ .

Thus we have $\hat{\Sigma}^{-1} = \Sigma^{-1} + O_p(\epsilon)\Sigma^{-1}E\Sigma^{-1} + o_p(\epsilon)$for large $n$

Using the fact that $\|\Sigma^{-1}\|_2 < \infty$ and $\Delta \sim N_{PT}(\Delta, \Sigma(\frac{1}{n_1} + \frac{1}{n_2}))$ ■

**Theorem 2.5.3.** Using assumptions A(1) to A(4) ,$\sum_{t=1}^{T} P_t = \tilde{P}_T = o(n)$ and $C_{\tilde{P}_T} \to C_\infty$ we show that $W(\hat{\delta}_{MLE}) \to 1 - \Phi(\frac{\sqrt{C_\infty}}{2})$

*Proof.* By Taylor Series expansion

$$\Phi(\frac{\hat{\Delta}^T\Sigma^{-1}\hat{\Delta}}{2\sqrt{\hat{\Delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\Delta}}})$$
$$= \Phi(\sqrt{\Delta^T\Sigma^{-1}\Delta}/2) + \big(\frac{\hat{\Delta}^T\Sigma^{-1}\hat{\Delta}}{2\sqrt{\hat{\Delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\Delta}}} - \sqrt{\Delta^T\Sigma^{-1}\Delta}/2\big)\phi(\sqrt{\Delta^T\Sigma^{-1}\Delta}/2) +$$
$$\frac{1}{2}(\frac{\hat{\Delta}^T\Sigma^{-1}\hat{\Delta}}{2\sqrt{\hat{\Delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\Delta}}} - \sqrt{\Delta^T\Sigma^{-1}\Delta}/2)^2\,(\sqrt{\Delta^T\Sigma^{-1}\Delta}/2)\;\phi(\sqrt{\Delta^{T*}\Sigma^{-1*}\Delta^*}/2)\big)$$

for some quantity $\Delta^{T*}\Sigma^{-1*}\Delta^* \in [\frac{\hat{\Delta}^T\Sigma^{-1}\hat{\Delta}}{2\sqrt{\hat{\Delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\Delta}}}, \sqrt{\Delta^T\Sigma^{-1}\Delta}/2]$

Now using lemma 2.5.2 and boundedness of $0 < \phi(.) \leq 1$

$$\Phi(\frac{\hat{\Delta}^T\Sigma^{-1}\hat{\Delta}}{2\sqrt{\hat{\Delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\Delta}}}) = \Phi(\lim_n \sqrt{\Delta^T\Sigma^{-1}\Delta}/2) + o_p(1)$$

■

**Lemma 2.5.4.**

$$\big(\frac{\hat{\Delta}^T\Sigma^{-1}\hat{\Delta}}{2\sqrt{\hat{\Delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\Delta}}} - \sqrt{\Delta^T\Sigma^{-1}\Delta}/2\big) = \frac{O_P(\frac{pT}{n})(\sqrt{\Delta^T\Sigma^{-1}\Delta})}{\sqrt{\Delta^T\Sigma^{-1}\Delta})(\sqrt{\Delta^T\Sigma^{-1}\Delta}) + O_P(\frac{pT}{n})}$$

*Proof.* From lemma 2.5.4, we get $\hat{\Sigma} = \Sigma + \epsilon E$ where $\epsilon = O_p(\sqrt{\frac{1}{pTn}})$ and $\|E\|_2 < \infty$ So by geometric series expansion is valid for large $n$ .Thus we have $\hat{\Sigma}^{-1} = \Sigma^{-1} + O_p(\epsilon)\Sigma^{-1}E\Sigma^{-1} + o_p(\epsilon)$ for large $n$. Using the fact that $\|\Sigma^{-1}\|_2 < \infty$ and $\Delta \sim N_{pT}(\Delta, (\frac{1}{n_1} + \frac{1}{n_2})\Sigma)$ ■

**Theorem 2.5.5.** Under assumption A(3) if $pT = o(n)$ and $\mathbb{C}_{pT} \to \mathbb{C}_\infty$ as $n \to \infty$, then $\mathbb{W}(\hat{\delta}_{MLE}) \to 1 - \Phi(\frac{\sqrt{C_\infty}}{2})$

*Proof.*

$$\Phi\Big(\frac{\hat{\Delta}^T\Sigma^{-1}\hat{\Delta}}{2\sqrt{\hat{\Delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\Delta}}}\Big)$$
$$= \Phi(\sqrt{\Delta^T\Sigma^{-1}\Delta}/2) + \Big(\frac{\hat{\Delta}^T\Sigma^{-1}\hat{\Delta}}{2\sqrt{\hat{\Delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\Delta}}}$$
$$- \sqrt{\Delta^T\Sigma^{-1}\Delta}/2\Big)\phi(\sqrt{\Delta^T\Sigma^{-1}\Delta}/2) +$$
$$\frac{1}{2}\Big(\frac{\hat{\Delta}^T\Sigma^{-1}\hat{\Delta}}{2\sqrt{\hat{\Delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\Delta}}}$$
$$- \sqrt{\Delta^T\Sigma^{-1}\Delta}/2)^2\,(\sqrt{\Delta^T\Sigma^{-1}\Delta}/2)\;\phi(\sqrt{\Delta^{T*}\Sigma^{-1*}\Delta^*}/2)\Big)$$

For some quantity $\Delta^{T*}\Sigma^{-1*}\Delta^* \in [\frac{\hat{\Delta}^T\Sigma^{-1}\hat{\Delta}}{2\sqrt{\hat{\Delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\Delta}}}, \sqrt{\Delta^T\Sigma^{-1}\Delta}/2]$

Now using lemma 2.5.4 and boundedness of $0 < \phi(.) \leq 1$

$$\Phi\Big(\frac{\hat{\Delta}^T\Sigma^{-1}\hat{\Delta}}{2\sqrt{\hat{\Delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\Delta}}}\Big) = \Phi(\lim_{n\to\infty}\sqrt{\Delta^T\Sigma^{-1}\Delta}/2) + o_p(1)$$

Results follow from Lemma 2.5.4. ■

With the application to neuroimaging, below we investigate the properties of the misclassification error rate whenever $pT >> n$. If we further assume a known covariance matrix $\Sigma$, with unknown mean estimates of the LDA classifier obtained using MLE, the following theorem provides the necessary motivation to modify the current method.

**Theorem 2.5.6.** Under the known true covariance $\Sigma$, let us define the LDA classifier,

$$\hat{\delta}_\mu(X) = (X - \hat{\mu})'\Sigma^{-1}(\hat{\mu}_1 - \hat{\mu}_2) \tag{2.18}$$

where $\hat{\mu}_1$,$\hat{\mu}_2$ are the MLE estimates and $\hat{\mu} = \frac{\hat{\mu}_1+\hat{\mu}_2}{2}$. Let us assume the high dimensional setup $pT/n \to \infty$ and let $\mathbb{C}_{pT} \to \mathbb{C}_\infty$ where $0 < \mathbb{C}_\infty < \infty$ then,

(1) For $n_1 \neq n_2$ but $n_k > n/4$ for $k \in \{1, 2\}$

(i) If $\frac{\sqrt{\mathbb{C}_{pT}}}{pT/n} \to c$ for $c > 0$, we get $\mathbb{W}(\hat{\delta}_\mu) \to 0$ but $\frac{\mathbb{W}(\hat{\delta}_\mu)}{1-\Phi(\frac{\sqrt{\mathbb{C}_{pT}}}{2})} \to \infty$

(ii) If $\frac{\sqrt{\mathbb{C}_{pT}}}{pT/n} \to 0$, we get $\mathbb{W}(\hat{\delta}_\mu) \to \frac{1}{2}$

(2) For $n_1 = n_2$,

(i) If $\frac{\sqrt{\mathbb{C}_{pT}}}{pT/n} \to \infty$, we get $\mathbb{W}(\hat{\delta}_\mu) \to 0$ but $\frac{\mathbb{W}(\hat{\delta}_\mu)}{1-\Phi(\frac{\sqrt{\mathbb{C}_{pT}}}{2})} \to \infty$

(ii) If $\frac{\sqrt{\mathbb{C}_{pT}}}{pT/n} \to c$ for $c > 0$, we get $\mathbb{W}(\hat{\delta}_\mu) \to 1 - \Phi(\frac{c}{4})$, which is a constant in $(0, \frac{1}{2})$

(iii) If $\frac{\sqrt{\mathbb{C}_{pT}}}{pT/n} \to 0$, we get $\mathbb{W}(\hat{\delta}_\mu) \to \frac{1}{2}$

*Proof.* We will prove that

$$\Phi(\frac{\hat{\Delta}^T\Sigma^{-1}\hat{\Delta}}{2}) = \Phi(\frac{\Delta^T\Sigma^{-1}\Delta)(1+O_p(\frac{1}{\sqrt{npT}}) + \frac{pT}{n}(n_1-n_2)(1+O_p(\frac{1}{\sqrt{npT}}))}{2\sqrt{\Delta^T\Sigma^{-1}\Delta)(1+O_p(\frac{1}{\sqrt{npT}}) + \frac{pT}{n}(n_1+n_2)(1+O_p(\frac{1}{\sqrt{npT}}))}})$$

The key step here is, $(\mu_1-\mu)^T\hat{\Sigma}^{-1}\hat{\Delta} = \Delta^T\Sigma^{-1}\Delta(1+O_p(\frac{1}{\sqrt{npT}})) + \frac{p}{n_1n_2}(n_1-n_2)(1+O_p(\frac{1}{\sqrt{npT}}))$.

$$(\mu_1-\mu)^T\hat{\Sigma}^{-1}\hat{\Delta} = \frac{1}{2}[(\Delta\hat{\Sigma}^{-1}\Delta)+(\hat{\mu_1}-\mu_1)^T\hat{\Sigma}^{-1}(\hat{\mu_1}-\mu_1)-(\hat{\mu_2}-\mu_2)^T\hat{\Sigma}^{-1}(\hat{\mu_2}-\mu_2)-2\Delta^T\hat{\Sigma}^{-1}(\hat{\mu_2}-\mu_2)]$$

from the previous results we have, $\hat{\Sigma}^{-1} = \Sigma^{-1}(1+O_p(\frac{1}{\sqrt{npT}}))$

■

In section 2.1, we have established that a variety of approaches have been investigated to overcome this singularity. Fan & Lv (2010) furthered emphasized on how the sparsity assumption and estimation may lead to optimal misclassification rates. Similar to the spatial approach adopted

by Yingjie & Maiti (2019), we expand methods to the spatio-temporal setup and incorporate non-separability into the following theoretical results.

## 2.6 Penalized Linear Discriminant Analysis (pLDA)

In most imaging studies with relatively higher resolution information, the number of spatial points (pixels and/or voxels) is almost exponentially higher than number of individual subjects for classification. Thus $pT > n$ is a rather unrealistic scenario. The LDA classifier under investigation too renders itself unsuitable in situations where acquisition of data results in $pT >> n$, as $\hat{\Sigma}$ is singular. As for the misclassification rate, we can also establish that under the setup where $pT/n \to \infty$ that the misclassification rate would be equivalent to random chance even when the true covariance is known due to the inconsistent accumulation of variance of the estimates of $\hat{\mu}_k$ for $k \in \{1, 2\}$. Therefore there is a need to develop methods that can better accommodate a much more likely scenario of $pT > n$. Although, based on setup we are able to reduce the number of estimates to $s + 2 << n$ dimensions where a solution maybe obtained using MLE, the nature of these estimates are unstable. Therefore we proceed by proposing a penalized LDA. To describe the method, we add to the notation used in section 2.4 as follows. Let $\Delta = \mu_1 - \mu_2 = (\Delta_1, ..., \Delta_{pT})'$ be a $pT \times 1$ dimensional vector which is the difference of the mean effect between classes $\mathcal{C}_1$ and $\mathcal{C}_2$. We define the signal set $\mathbb{S} = \{\nu : \Delta_\nu \neq 0\}$, denoting the true non-zero differences between classes. Further, let $s$ denote the cardinality of $\mathbb{S}$ i.e. $s = |\mathbb{S}|$ and we assume that $s < n << pT$. To formalize vector and matrix representation of the model in order to easily incorporate the penalty term, for $Y_{k,i} \sim \mathbf{N}_{pT}(\mu_k, \Sigma(\theta))$, we define $Z_i = Y_{1,i} - \bar{Y}$ for $i = 1, 2, ..., n_1$ and $Z_i = Y_{2,i} - \bar{Y}$ for $i = n_1 + 1, n_1 + 2, ..., n - 1$ where $\bar{Y} = \sum_{k=1,2} \sum_{i=1}^{n_k} Y_{k,i}/2n$ and $n = n_1 + n_2$. Let $\tau_1 = \frac{n_1}{n}$ and

$\tau_2 = \frac{n_2}{n}$. Thus,

$$Z_i \sim \begin{cases} \mathbf{N}_{pT}(-\tau_2\,\Delta, \frac{n-1}{n}\Sigma(\theta)), & \text{for } i = 1, 2, .., n_1 \\ \mathbf{N}_{pT}(\tau_1\,\Delta, \frac{n-1}{n}\Sigma(\theta)), & \text{for } i = n_1 + 1, 2, .., n-1 \end{cases}$$

and $\text{Cov}(Z_i, Z_j) = \frac{1}{n}\Sigma(\theta))$ for $i \neq j$. Alternatively,

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_{n-1} \end{pmatrix} \sim \mathbf{N}\Bigg( \underbrace{\begin{pmatrix} -\tau_2\mathbf{1}_{n_1} \\ \\ \tau_1\mathbf{1}_{n_2-1} \end{pmatrix} \otimes \mathbf{I}_{pT}}_{\tilde{\mathbf{X}}} \underbrace{\Delta}_{\beta}, \underbrace{(\mathbf{I}_{n-1} - \frac{1}{n}\mathbf{J}_{n-1}) \otimes \Sigma(\theta)\Bigg)}_{\tilde{\Sigma}(\theta)} \tag{2.19}$$

where $\mathbf{I}_b$ signifies a $b \times b$ identity matrix and $\mathbf{J}_b$ signifies a $b \times b$ matrix of 1s while $\mathbf{1}_b$ denotes a $b \times 1$ dimensional vector. We also have, $\tilde{\Sigma}^{-1}(\theta) = (\mathbf{I}_{n-1} + \mathbf{J}_{n-1}) \otimes \Sigma^{-1}(\theta)$ and $\mathbf{X} = \begin{pmatrix} -\tau_2\mathbf{1}_{n_1} \\ \\ \tau_1\mathbf{1}_{n_2-1} \end{pmatrix}$ such that $\tilde{\mathbf{X}} = \mathbf{X} \otimes \mathbf{I}_{pT}$.

We now express the joint log-likelihood of Z as,

$$\mathcal{L}(\beta, \theta; \mathbf{Z}) = -\frac{npT}{2}log(2\pi) - \frac{1}{2}(\mathbf{Z} - \tilde{\mathbf{X}}\beta)^T\tilde{\Sigma}^{-1}(\mathbf{Z} - \tilde{\mathbf{X}}\beta) - \frac{1}{2}log|det(\tilde{\Sigma})| \tag{2.20}$$

It is clear that we seek to obtain a solution for a sparse true $\beta_0 = (\beta'_{1,0}, \beta_{0,2})' = (\beta'_{1,0}, 0)'$ where $\beta_{1,0}$ is an $s$ dimensional vector, $\beta_{2,0}$ is a $p - s$ dimensional vector and $\beta_0$ is $pT \times 1$. In this high-dimensional setup, we get solutions for estimating $\theta$ and $\beta$ by solving the penalized log-likelihood given by,

$$\mathcal{Q}(\beta, \theta; \mathbf{Z}) = \mathcal{L}(\beta, \theta; \mathbf{Z}) - n\sum_{t=1}^{T}\sum_{u=1}^{p} p_{\lambda_{t,n}}(|\beta_u|) \tag{2.21}$$

Here $p_{\lambda_{t,n}}$ denotes the penalty function $p$ with its corresponding tuning parameter $\lambda_{t,n}$ depend-

ing on $n$. To obtain, well defined properties of the sparse estimator we consider the smoothly clipped absolute deviation (SCAD, Fan & Li (2001)) penalty function given by,

$$p_{\lambda_n}(\beta) \sim \begin{cases} \lambda_{t,n}|\beta|, & \text{if } |\beta| < \lambda_{t,n} \\ -\frac{\beta^2 - 2\alpha\lambda_{t,n}\beta + {\lambda_{t,n}}^2}{2(\alpha-1)}, & \text{if } \lambda_{t,n} < |\beta| < \alpha\lambda_{t,n} \\ \frac{(\alpha+1){\lambda_{t,n}}^2}{2}, & \text{if } |\beta| > \alpha\lambda_{t,n} \end{cases}$$

### 2.6.1 Across sample independence

Notice that samples are not uncorrelated due to covariance matrix $\underbrace{(\mathbf{I}_{n-1} - \frac{1}{n}\mathbf{J}_{n-1})}_{HH^T} \otimes \Sigma(\theta)$ but we can make a linear transformation to data $\mathbf{Z}'_n = (H \otimes \mathbf{I}_{PT})\mathbf{Z}_n$ which ensures no correlation across sample in $Z$ thus independence in Gaussian case. Take $H_n = (\mathbf{I}_{n-1} - \frac{1}{\sqrt{n}+1}\mathbf{J}_{(n-1)})$ and resulting design matrix would be $X' = H_n X = (\mathbf{I}_{n-1} - \frac{1}{\sqrt{n}+1}\mathbf{J}_{(n-1)} \begin{pmatrix} -\tau_2 \mathbf{1}_{n_1} \\ \tau_1 \mathbf{1}_{n_2-1} \end{pmatrix} = \begin{pmatrix} (-\tau_2 - \frac{\tau_1}{\sqrt{n}+1})\mathbf{1}_{n_1} \\ (\tau_1 - \frac{\tau_1}{\sqrt{n}+1})\mathbf{1}_{n_2-1} \end{pmatrix}$
Note that $\tilde{X}^T\tilde{\Sigma}^{-1}\tilde{X} = \tilde{X'}^T\tilde{\Sigma'}^{-1}\tilde{X'}$

Henceforth we refer to transformed likelihood equation as our original equation and with the abuse of notation, we assign the notation of original variable to transformed variable. We now express the joint log-likelihood of Z as,

$$\mathcal{L}(\beta, \theta; \mathbf{Z}) = -\frac{npT}{2}log(2\pi) - \frac{1}{2}(\mathbf{Z} - \tilde{\mathbf{X}}\beta)^T\tilde{\Sigma}^{-1}(\mathbf{Z} - \tilde{\mathbf{X}}\beta) - \frac{1}{2}log|det(\tilde{\Sigma})| \tag{2.22}$$

It is clear that since we seek to obtain a solution for a sparse true $\beta_0 = (\beta'_{1,0}, 0)'$ where $\beta_{1,0}$ is an $s$ dimensional vector and $\beta_0$ is $pT \times 1$ in this high-dimensional setup, we get solutions for estimating $\theta$ and $\beta$ by solving the penalized log-likelihood given by,

$$\mathcal{Q}(\beta, \theta; \mathbf{Z}) = \mathcal{L}(\beta, \theta; \mathbf{Z}) - n\sum_{t=1}^{T}\sum_{u=1}^{p} p_{\lambda_t}(|\beta_u|) \tag{2.23}$$

Here $p_\lambda$ denotes the penalty function $p$ with its corresponding tuning parameter $\lambda_n$ depending on $n$. To obtain, well defined properties of the sparse estimator we consider the smoothly clipped absolute deviation (SCAD, Fan & Li (2001)) penalty function given by,

$$p_{\lambda_n}(\beta) \sim \begin{cases} \lambda_n|\beta|, & \text{if } |\beta| < \lambda_n \\ -\frac{|\beta|^2 - 2\alpha\lambda_n|\beta| + {\lambda_n}^2}{2(\alpha-1)}, & \text{if } \lambda_n < |\beta| < \alpha\lambda_n \\ \frac{(\alpha+1){\lambda_n}^2}{2}, & \text{if } |\beta| > \alpha\lambda_n \end{cases}$$

### 2.6.2 REML estimation

Without loss of generality we can expand

$$\beta_{0\,PT\times 1} = (\beta_{0\,1,P_1\times 1}, \beta_{0\,2,P_2\times 1}, .\beta_{0\,t,P_t\times 1}, ..\beta_{0\,T,P_T\times 1})^T$$

For each $t$ we can further expand $\beta^T_{0\,t,P_t\times 1} = (\beta_{0\,1,t,s_t\times 1}, \mathbf{0}_{P_T - s_t\times 1})^T$

Notice that $X\otimes\mathbb{I}_{PT}\beta = X\otimes\oplus_{t=1}^T \begin{pmatrix} \mathbb{I}_{s_t} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} (\beta_{0\,1,1,s_t\times 1}, \mathbf{0}_{P_t - s_t\times 1} ..\beta_{0\,1,t,s_t\times 1}, \mathbf{0}_{P_t - s_t\times 1} ..\beta_{0\,1,T,s_T\times 1}, \mathbf{0}_{P_T - s_T\times 1})^T$

$$= X \otimes \underbrace{\oplus_{t=1}^T \begin{pmatrix} \mathbb{I}_{s_t} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}}_{Q_{PT\times PT}} \beta$$

In the context of this result let us redefine $\tilde{X} = X \otimes Q$ Clearly rank of $\tilde{X} = nPT - \sum_{t=1}^T s_t$

Suppose, we obtain a orthogonal matrix $B_n$ such that

$B_n^T B_n = \oplus_{t=1}^T diag \begin{pmatrix} \mathbf{I_{n(P_t - s_t)}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$ and $B_n^T B_n = \mathbf{I} - \tilde{X}(\tilde{X}^T\tilde{X})^-\tilde{X}^T$

REML estimates are defined by

$$\hat{\theta}_{n,REML} = \arg\min_\theta \; -\log\det(\tilde{\Sigma}(\theta)) - \log\det(\tilde{X}^T\tilde{\Sigma}(\theta)^{-1}\tilde{X}) - \frac{1}{2}\boldsymbol{Z}^T\Pi(\theta)\boldsymbol{Z}$$

Let us define oracle estimator $\hat{\boldsymbol{\beta}}_t^{orc}$ when it is set of zeros in $\beta_{t,0}$ is already known. Under that setup, $\hat{\boldsymbol{\beta}}_t^{orc}|_{\hat{\theta}} = (\tilde{X}_{1,t}^T\tilde{\Sigma}^{-1}(\hat{\theta})\tilde{X}_{1,t})^-\tilde{X}_{1,t}^T\tilde{\Sigma}^{-1}(\hat{\theta})\mathbf{Z}_n$ where $\tilde{X}_{1,t} = \mathbb{I}_{s_t}$ This leads to equation

$$trace\left(\Pi(\hat{\theta}_{n,REML})\frac{\partial\tilde{\Sigma}(\hat{\theta}_{n,REML})}{\partial\theta}\right) + \mathbf{Z}^T\frac{\partial\Pi(\hat{\theta}_{n,REML})}{\partial\theta}\mathbf{Z} = \mathbf{0} \tag{2.24}$$

where $\tilde{\Pi}(\theta) = \tilde{\Sigma}^{-1}(\theta) - \tilde{\Sigma}^{-1}(\theta)\tilde{X}_1(\tilde{X}_1^T\tilde{\Sigma}^{-1}(\theta)\tilde{X}_1)^-\tilde{X}_1^T\tilde{\Sigma}^{-1}(\theta)$

The attractive features of REML are that equation 5.4 is unbiased and the covariance estimate $\hat{\theta}_{REML}$ and mean estimate $\hat{\beta}$ are asymptotically independent.

**Theorem 2.6.1.** Oracle REML Cressie & Lahiri (1996)

$\|\hat{\theta}_{n,REML} - \theta_0\|_2 = O_p(\sqrt{\frac{1}{\sum_{t=1}^T P_t(n)}})$

*Proof.* From the result, $Pr(\hat{\beta}_t|_{\hat{\theta}} = \hat{\beta}_t^{orc}|_{\hat{\theta}}) \to 1$ and from the assumption that $\{j : \hat{\beta}_{t,j}^{orc} = 0\} = \{j : \beta_{t,j} = 0\}$, we know the indices or position where $\beta_0$ is zero. therefore we are able to identify with probability one about linear subspace of design matrix generated by $\tilde{X}_{1,t}$ needed to construct projected covariance matrix $\tilde{\Pi}$ .So, with probability converging to 1, we get result as stated in Cressie & Lahiri (1996) ■

### 2.6.3 Validation of REML assumptions

From Cressie & Lahiri (1996) ,we observe that assumption A1-A4 are sufficient to guarantee convergence of covariance parameter $\hat{\theta}$. In this section we prove those assumption hold for Matern covariance class under suitable chosen parameter space.

**Lemma 2.6.2.** For every isotropic , stationary function $\gamma(h,t)$ the matrix $\Sigma$ is positive definite with entries $\Sigma_{i,j} = \gamma(s_i - s_j, t_i - t_j)$ for some fixed $P$ spatial points $s_i, s_j \in \mathcal{S}$ and $T$ time points $t_i, t_j \in \mathcal{T}$

*Proof.* For any vector $\underset{\sim}{a}$; $\underset{\sim}{a}^T\Sigma\underset{\sim}{a} = Var(a^T\underset{\sim}{X})$ here $covar(X_i, X_j) = \gamma(s_i - s_j, t_i - t_j)$ ■

conditions C4-C6 is sufficient for assumptions A4 which guarantees convergence of REML estimators even in irregular time domain. Since we assume our covariance follows Matern class which is isotropic and second order stationary we can establish following lemma:

**Lemma 2.6.3.** Conditions A4.1-3 , A4.5 and A4.6 are satisfied if the item 1 , item 2 and item 3-4 holds for any isotropic and second order stationary process.

For any $\theta \in K$ compact subset

- $\lim_n \limsup_{s\in\mathcal{S}_t, t\leq T_n} \sum_{s'\in\mathcal{S}_t, t'\leq T_n} |\gamma_{(.)}(s-s', t-t'; \theta)| < \zeta_{1,K}$ here $\gamma_{(.)} = \gamma_0, \gamma_i\gamma_{i,j}$ denoting covariance, first and second derivative of covariance respectively

- $\lim_n \limsup_{s\in\mathcal{S}_t, t\leq T_n} \sum_{s'\neq s\in\mathcal{S}_t, t\neq t\leq T_n} |\gamma_0(s-s', t-t'; \theta)| < \zeta_{2,K}\gamma_0(\mathbf{0}, 0; \theta)$

- for all $i$ $\lim_n \limsup_{s\in\mathcal{S}_t, t\leq T_n} \sum_{s'\in\mathcal{S}_t, t\leq T_n} \gamma_i^2(s-s', t-t'; \theta) > \zeta_{2,K}$

- for all $i$ and $\mathcal{N}(s) = \{s' : s + s' \in \mathcal{S}_t\}$ and $\mathcal{N}(t) = \{t' : t + t' \in \mathcal{S}_t\}$

$$\lim_n \sum_{s\in\mathcal{S}_t, t\leq T_n} |\gamma_i(s,t;\theta)| \sum_{s'\neq s\in\mathcal{S}_{1t}, t'\neq t\in T_{1n}} |\gamma_i(s-s', t-t'; \theta)|$$
$$< \zeta_{2,K} \lim_n \sum_{s'\in\mathcal{N}(s), t'\in\mathcal{N}(t)} \gamma_i^2(s', t'; \theta) \text{ for each fixed } s \in \mathcal{S}_{1,t}, t \leq T_{1,n}$$

*Proof.* Since closed form of $\gamma_{(.)}$ is not available for all values for separability parameter $\epsilon$ , we use the argument that eigenvalues of $\Sigma$ and $\frac{\partial\Sigma}{\partial\theta_i}$ are continuous in $\epsilon$ due to continuity of $f()$ and $\frac{\partial f}{\partial\theta_i}$ as function of $\epsilon$. Our strategy is to prove all the results for $\epsilon = 0, 1$ for which closed form are available. Also notice that spectral density $f()$ is continuously differentiable in $\epsilon$.

$$\gamma_0(\|s\|, t) = \begin{cases} \frac{\eta\pi^{\frac{d}{2}}\Gamma(\nu+\frac{d}{2})(\alpha\|s\|)^\nu\mathcal{K}_\nu(\alpha\|s\|)}{2^{\nu-1}\alpha^{2\nu}} \frac{\pi^{\frac{1}{2}}\Gamma(\nu+\frac{1}{2})(\beta|t|)^\nu\mathcal{K}_\nu(\beta|t|)}{2^{\nu-1}\beta^{2\nu}} & \epsilon = 1 \\ \frac{\eta\pi^{\frac{d+1}{2}}\Gamma(\nu+\frac{d+1}{2})(\sqrt{\alpha^2\|s\|^2+\beta^2t^2})^\nu\mathcal{K}_\nu(\sqrt{\alpha^2\|s\|^2+\beta\|^2t^2})}{2^{\nu-1}(\alpha^2+\beta^2)^\nu} & \epsilon = 0 \end{cases}$$

Let $k \geq 0$ be an integer define the set $E_k = s' : \|s' - s\| \leq k + 1/s' : \|s' - s\| \leq k$ and $N_k = |E_k|$ number of points in $E_k$. To each point $s' \in E_k$ we associate a disjoint $\|\|_2$ ball of volume $2\frac{\pi^{\frac{d}{2}}(\frac{a}{2})^d}{\Gamma(1+\frac{d}{2})}$ , so total space occupied is $N_k 2\frac{\pi^{\frac{d}{2}}(\frac{a}{2})^d}{\Gamma(1+\frac{d}{2})}$, also volume of $E_k = 2\frac{\pi^{\frac{d}{2}}\left((k+1)^d - k^d\right)}{\Gamma(1+\frac{d}{2})} \leq 2\frac{\pi^{\frac{d}{2}}\left(d(k+1)^{d-1}\right)}{\Gamma(1+\frac{d}{2})}$ therefore we obtain $N_k \leq \frac{d(k+1)^{d-1}2^d}{a^d}$ and since non decreasing behavior as a function of distance between arguments. $\sup_{s' \in E_k} \gamma_0(s - s', t - t') \leq \gamma_0(k, t - t')$

$$\lim_n \limsup_{s \in \mathcal{S}_t, t \leq T_n} \sum_{s' \in \mathcal{S}_t, t' \leq T_n} |\gamma_0(s - s', t - t'; \theta)| = \lim_n \limsup_{s \in \mathcal{S}_t, t \leq T_n} \sum_{s' \in \mathcal{S}_t, t' \leq T_n} \gamma_0(s - s', t - t'; \theta)$$
$$\leq \sum_{k \in \mathbb{Z}^+, k_2 \in \mathbb{Z}^+} \frac{d(k+1)^{d-1}2^d}{a^d} \gamma_0(k, bk_2; \theta) = \frac{1}{b} \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} \frac{d(s+1)^{d-1}2^d}{a^d} \gamma_0(s, t) ds dt$$
$$\leq C_K \frac{1}{b} \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} \frac{d(s+1)^{d-1}2^d}{a^d}(s, t) ds dt = O(1)$$

due to finite moment of modified Bessel function of second kind

- From the property of uniform integrability ,

■

#### 2.6.3.1 Tapered REML

Suppose we taper each spatial matrix with fixed tapered range $r_t = \frac{K_t}{\sqrt{P_t}}$ such that constant $K_t$ is to be determined by cross validation.

We solve modified REML equation

$$trace\left(\Pi(\hat{\theta}_{n,REML,tapered}))_{tap} \frac{\partial \tilde{\Sigma}(\hat{\theta}_{n,REML})}{\partial \theta}\right) + \mathbf{Z}^T \frac{\partial \Pi(\hat{\theta}_{n,REML,tapered})_{tap}}{\partial \theta} \mathbf{Z} = \mathbf{0} \tag{2.25}$$

Here $\Pi_{tap} = \Sigma \circ \tilde{K}_{Tap}{}^{-1} - \Sigma \circ \tilde{K}_{Tap}{}^{-1} \tilde{X}_1 (\tilde{X}_1^T \Sigma \circ \tilde{K}_{Tap}{}^{-1} \tilde{X}_1)^- \tilde{X}_1^T \Sigma \circ \tilde{K}_{Tap}{}^{-1}$

In order to show the convergence of tapered REML estimater i.e. $\|\hat{\theta}_{n,REML,tapered} - \theta_0\|_2 = O_p(\sqrt{\frac{1}{\sum_{t=1}^T P_t(n)}})$ we ensure that even after tapering all conditions as mentioned in Cressie & Lahiri

(1996) remains intact. Hence we need to prove the following lemma which assures that tapered in not "far" from original covarinace matrix

**Lemma 2.6.4.** Assuming general condition 13 holds, we have the results that for each t,

- $\|\Sigma_t - \Sigma_{t,Taper}\|_1 = O_p(\frac{1}{\sqrt{P_t}})$
- $\|\Sigma_{k,t} - \Sigma_{k,t,Taper}\|_1 = O_p(\frac{1}{\sqrt{P_t}})$
- $\|\Sigma_{k,j,t} - \Sigma_{k,j,t,Taper}\|_1 = O_p(\frac{1}{\sqrt{P_t}})$

*Proof.*

$$\|\Sigma_t - \Sigma_{t,Taper}\|_1 = \sum_{t=1}^{T}\sum_{1}^{p_t} |\gamma_{taper}(s,t:\theta)|$$

$$= \max_i \sum_{h_{ij}\leq w_{p_t}} |\gamma_{taper}(s,t:\theta)| + \max_i \sum_{h_{ij}\geq w_{p_t}} |\gamma_{taper}(s,t:\theta)|$$

$$\leq K_t\rho \sum_{m=\rfloor \frac{w(p_t)}{\delta} \lfloor} \sum_{j\in B_m^i} \leq 3\frac{K_t\rho}{w(p_t)} \int_{w_p}^{\infty} x^d |\gamma(x:\theta)| dx$$

■

Therefore based on equation 2.23 in order to estimate $\hat{\beta}$ and $\hat{\theta}$, we obtain the solutions of derivative of the penalized likelihoods with respect to each of the unknown parameters and iteratively solve until convergence is obtained.

### 2.6.4 Regularity conditions for penalty

In order to demonstrate properties of the subject level images over time, we introduce $\beta_{t,0,j}$ that is the unknown true parameters of dimension $p \times 1$ for each fixed time point $t = \{1, 2, .., T\}$. Analogously under the sparsity assumption, we denote $s_t$ to determine the number of non-zero components within a specific time point. Therefore, $\sum_{t=1}^{T} s_t = s$. The methods below without loss

of generality, as voxelwise analysis that involves a subject wise registration on a single template space are performed, that each image contains the same number spatial sites $p$. We further assume that for every subject the multiple images were acquired over time with a total number of $T$ time points. Clearly, for separable cases this provides the ease of the Kronecker product. $\Sigma(\theta)$ is assumed to be second order stationary, isotropic and twice differential over all dimensions of space and time for $\theta \in \Xi$ over all $(s,t) \in D \times \mathcal{T}$ where $\Xi$ is the parametric space of $\theta$. In general, let us consider any covariance matrix $\Sigma(\theta)$ for $\theta \in \Xi$ constructed by a covariance function $\gamma(\theta)$ where the true parameter $\theta$ is given by $\theta_0$. Also note that the derivative of $\Sigma(\theta)$ w.r.t $\theta_m$ is denoted by $\Sigma^m(\theta) = \frac{\partial}{\partial\theta_m}\Sigma(\theta)$ and second derivative is denoted by $\Sigma^{mm'}(\theta) = \frac{\partial^2\Sigma(\theta)}{\partial\theta_m\partial\theta_{m'}}$ where $\theta = (\theta_1, ..\theta_m, ..\theta_r)'$. Listed below are conditions under which consistent estimates of the pLDA are obtained.

(A3) $a_{t,n} = \max\limits_{1\leq j\leq p}\{p'_{\lambda_{t,n}}(|\beta_{t,0,j}|), \beta_{0,j} \neq 0\} = O(\sqrt{n})$

(A4) $b_{t,n} = \max\limits_{1\leq j\leq p}\{p''_{\lambda_{t,n}}(|\beta_{t,0,j}|), \beta_{0,j} \neq 0\} = o(1)$

(A5) $\frac{\sqrt{n}\lambda_{t,n}}{\sqrt{s}} \to \infty$

(A6) $\lambda_{t,n} = o(1)$ as $n \to \infty$.

(A7) $\frac{s^4}{n} \to 0$ as $n \to \infty$.

(A8) $\min\limits_{1\leq i\leq s} \frac{|\beta_{t,0,i}|}{\lambda_{t,n}} \to 0$ as $n \to \infty$.

(A9) $\lim\limits_{n\to\infty} \liminf\limits_{\theta\to 0^+} p'_{\lambda_{t,n}}(|\theta|) > 0$ as $n \to \infty$.

(A10) $\frac{\sum_1^T s_t \sum_1^T d_{p_t}}{nC_{pT}} \to 0$ where $d_p = max_{i\leq s_t} \sum_{k=s_t+1}^{p_t} \sigma_{k,t}^2$

We can rearrange $\beta_0$ and re-express it as,

$$\beta_{0\,pT\times 1} = (\beta_{0\,1,p\times 1}, \beta_{0\,2,p\times 1}, ..., \beta_{t,0,p\times 1}, ..\beta_{T,0,p\times 1})^T$$

For all of the methods shown below the only constraint for identifiability is that $\sum_{t=1}^{T} s_t = o(n)$. Considering $\lambda_{t,n} = \lambda_n$ is equal for each time group, below are results under these special cases.

**Lemma 2.6.5.** Assuming $\|\hat{\theta} - \theta\|_2 = O_p(\sqrt{\frac{1}{PTn}})$ , we want to prove that for each $t = 1, 2..., T$ $\|\hat{\beta}_t - \beta_{t,0}\|_2 = O_p(\sqrt{\frac{s_t}{n}})$ and $\hat{\beta}_{t\,2} = 0_{P\times 1}$ with probability close to 1 where $\hat{\beta}_t|_{\hat{\theta}} = \arg\max_{\beta_t} \mathbf{Q}(\boldsymbol{\beta}, \hat{\theta}; \boldsymbol{Z})$

*Proof.* For each fixed $t$ the proof consists of two steps

In first part we show that $\|\hat{\beta}_t - \beta_{t,0}\|_2 = O_p(\sqrt{\frac{s_t}{n}})$ . In the next part we show that $\hat{\beta}_{t\,2} = 0_{P\times 1}$ with probability close to 1 . Thus the proof is in two parts :

Without loss of generality, $\beta_{0,t} = (\beta_{0,t,1}, \mathbf{0})^T$ and lets assume that an oracle has already informed us about the positions of zero, thus we can construct a oracle estimator which already have zeros in the positions of

$$\hat{\beta}_t^{orc}|_{\hat{\theta}} = \underset{\beta_{t,1}|\beta_{t,1}^c=\mathbf{0}}{\arg\max}\ \mathbf{L}(\boldsymbol{\beta}, \hat{\theta}; \boldsymbol{Z})$$

By this construction we are only searching for $\hat{\beta}_t^{orc}$ in the space of real vectors whose non zero element corresponds to non zero part of true $\beta_{t,0}$.

In the second part we show that for any maximizer of penalized likelihood $\hat{\beta}_t|_{\hat{\theta}} = \arg\max_{\beta} \mathbf{Q}(\boldsymbol{\beta}, \hat{\theta}; \boldsymbol{Z})$

$$Pr(\hat{\beta}_t|_{\hat{\theta}} = \hat{\beta}_t^{orc}|_{\hat{\theta}}) \to 1$$

Furthermore if Likelihood is concave like in our Gaussian case , the maximizer $\hat{\beta}_t$ is unique. To prove second part we use the stationary condition of KKT conditions refereed in Kwon & Kim (2012)

The proof of first part

**Theorem 2.6.6.** Assuming $\|\hat{\theta} - \theta\|_2 = O_p(\sqrt{\frac{1}{PTn}})$ , we want to prove that for each $t = 1, 2..., T$ $\|\hat{\beta}_t^{orc}|_{\hat{\theta}} - \beta_{t,0}\|_2 = O_p(\sqrt{\frac{s_t}{n}})$

Denote vector $u_t \in \mathbb{R}^{p_n}$ with entries 1 corresponding to index j such that $\beta_{t,j} \neq 0$ and 0 elsewhere, but without loss of generality , we denote $u_t$ as $u$.Similarly, $\tilde{X}_t$ denotes the colomns of design matrix $\tilde{X}$ corresponding to $\beta_{t,j} \neq 0$. Hence we get identity that $\tilde{X}_t u_t = \tilde{X} \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbb{I}_{s_t \times s_t} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} u$

It is sufficient to show that for any $\epsilon > 0$ ,$\sum_{t=1}^{T} s_t = s$, $\xi_{n,t} = O_p(\sqrt{\frac{s_t}{n}})$ , $\xi_n = O_p(\sqrt{\frac{s}{n}})$ and for some $C > 0$ under the condition that for $r \in \{1, 2, ....T\} - \{t\}$ we have $\|\hat{\beta}_r - \beta_r\|_2 = O_p(\sqrt{\frac{s_r}{n}})$

$$Pr\big( \sup_{\|u\|_2 = C} \mathbf{L}(\beta_{t,0} + u\xi_{n,t}, \hat{\theta}) - \mathbf{L}(\beta_{t,0}, \hat{\theta}) \ \big) > 1 - \epsilon$$

or equivalently we can prove that

$$Pr\big( \sup_{\|u\|_2 = C} \mathbf{L}(\beta_{0,t} + u\xi_{n}, \hat{\theta}) - \mathbf{L}(\beta_{0,t}, \hat{\theta}) \ \big) > 1 - \epsilon$$

For term (1) we get using eigenvalue inequality result that $\psi_i$ is $i$ th eigenvalue

*Proof.*

$$\underbrace{\frac{1}{2} u_t^T \tilde{X}_t^T \tilde{\Sigma}(\hat{\theta})^{-1} \tilde{X}_t u_t \xi_n}_{1}$$

$$= \underbrace{\frac{1}{2} u^T \tilde{X}^T \tilde{\Sigma}(\theta_0)^{-1} \tilde{X}_t u_t \xi_n}_{1a} + \underbrace{\frac{1}{2} u_t^T \tilde{X}_t^T \sum_{r=1}^{q} u^T \tilde{X}^T \dot{\tilde{\Sigma}}(\theta^*)^r (\hat{\theta}_r - \theta_{0,r}) u^T \tilde{X}^T \xi_n}_{1b}$$

$$1(a) \leq O(\|u\|^2 \tilde{X_t}^T \tilde{X}_t \min_i \psi_i(\tilde{\Sigma}(\hat{\theta})^{-1})) \xi_n^2$$

$$= O(\min_i \psi_i(\tilde{\tilde{X}}_t^T (\mathbf{I}_{n-1} + \mathbf{J}_{n-1})) \tilde{X}_t) \min_i \psi_i(\tilde{\Sigma}(\hat{\theta})^{-1})) \xi_n^2$$

$$= O(n\xi_{n,t}^2) \text{by eigenvalue inequality}$$

Similarly using CS inequality$1(b) = O(n\xi_{n,t}^2 \eta_n)$

■

For term (2) we use Taylor expansion of $\tilde{\Sigma}(\hat{\theta})^{-1}$ around $\theta_0$ such that

$$\hat{\theta} - \theta_0 = v\eta_n$$

*Proof.* $-(Z - \tilde{X}\beta)^T\tilde{\Sigma}(\hat{\theta})^{-1}\tilde{X}_t u_t$

$$= \underbrace{-(Z - \tilde{X}\beta)^T\tilde{\Sigma}(\theta_0)^{-1}\tilde{X}u\xi_n}_{2a} - \underbrace{\sum_{r=1}^{q}(Z_t - \tilde{X}\beta_t)^T\dot{\tilde{\Sigma}}(\theta^*)^r(\hat{\theta}_r - \theta_{0,r})\tilde{X}u\xi_n}_{2b}$$

term (2a) using Markov inequality that for any r.v

$Z = O_p\sqrt{\mathbb{E}(Z^2)}$

$$-(Z - \tilde{X}\beta)^T\tilde{\Sigma}(\theta_0)^{-1}\tilde{X}_t u_t\xi_n = O_p(\sqrt{\min_i \psi_i(\tilde{\Sigma}(\theta_0)^{-1})\, trace(\tilde{\tilde{X}}^T\tilde{\tilde{X}})}\|u\|_2\xi_n)$$

$$= O_p(\sqrt{\min_i \psi_i(\Sigma(\theta_0)^{-1})\, trace(\tilde{X}_t^T\tilde{X}_t)\begin{pmatrix} -\tau_2\mathbf{1}_{n_1} \\ \\ \tau_1\mathbf{1}_{n_2-1}\end{pmatrix}^T(\mathbf{I}_{n-1} + \mathbf{J}_{n-1})\begin{pmatrix} -\tau_2\mathbf{1}_{n_1} \\ \\ \tau_1\mathbf{1}_{n_2-1}\end{pmatrix}}\|u\|_2\xi_n)$$

$$= O_p(\sqrt{\tfrac{n_1n_2}{n}s_t}\|u\|\xi_n) = O_p(\sqrt{ns_t}\|u\|\xi_n)$$ ■

Similarly term (2b)

$$\textstyle\sum_{r=1}^{q}(Z - \tilde{X}\beta)^T\tilde{\Sigma}(\theta^*)^r\tilde{X}u\xi_n$$

$$= O_p(\sqrt{\min_i \psi_i(\tilde{\Sigma}(\theta^*)^r\tilde{\Sigma}(\theta_0)\tilde{\Sigma}(\theta^*)^r)trace\tilde{X}^T\tilde{X})}\|u\|_2\|v\|_2\xi_n\eta_n)$$

$$= O_p(\sqrt{ns_t}\|u\|\xi_n\eta_n) = o_p(\sqrt{ns_t}\|u\|\xi_n)$$ ■

Here we observe that term 1 dominates all other term hence for appropriate constant $C$ the value above is negative, hence proved.

**Lemma 2.6.7.** $Pr(\min_{j\le q_n}|\hat{\beta}_{t,j}|_{\hat{\theta}}| > a\lambda_n) \to 1$

*Proof.*

$$\min_{j\le q_n}|\hat{\beta}_{t,j}|_{\hat{\theta}}| \le \min_{j\le q_n}|\beta_{0,t,j}| - \max_{j\le q_n}|\hat{\beta}_{t,j}|_{\hat{\theta}} - \beta_{0,t,j}|$$

Using result $\|\hat{\beta}_{t,j}|_{\hat{\theta}} - \beta_{0,t,j}\|_2 = O_p(\sqrt{\frac{s_t}{n}})$ and assumption A6 we establish the above result. ■

**Lemma 2.6.8.** $Pr(\max_{q_n+1\le j\le p_n} |\frac{\partial \mathbf{L}(\hat{\beta}^{orc}_{t,j}|_{\hat{\theta}},\hat{\theta})}{\partial \beta}| < n\lambda_n|_{\theta=\hat{\theta}} \text{ for all} t) \to 1$

*Proof.* $\mathbb{E}_{\beta,\theta}\frac{\partial \mathbf{L}(\beta_0,\theta=\hat{\theta})}{\partial \beta} = 0$ due to conditional expectation.Using lemma 9.1, we can prove that $\frac{\partial \mathbf{L}(\hat{\beta}^{orc}_{t,j}|_{\hat{\theta}},\hat{\theta})}{\partial \beta} = -(\boldsymbol{Z} - \tilde{X}\hat{\beta}^{orc}_{t,j}|_{\hat{\theta}})^T\tilde{\Sigma}^{-1}(\hat{\theta})\tilde{X}_{1,t}$

$= -(\boldsymbol{Z} - \tilde{X}\beta_0)^T\tilde{\Sigma}^{-1}(\hat{\theta})\tilde{X}_{1,t} + -(\hat{\beta}^{orc}_{t,j}|_{\hat{\theta}} - \beta_0)^T\tilde{X}^T$ By Gaussian concentraion inequality each of this quantity is $PTO(e^{-\lambda^2_{n,t}})$ Hence proved. ■

**Lemma 2.6.9.** $\|\hat{\beta} - \beta_0\| = O_p(\sqrt{\sum^T \frac{s_t}{n}})$ then $\|\hat{\theta} - \theta_0\| = O_p(\sqrt{\frac{1}{\sum_1^T P_t n}})$

*Proof.* We will show that for arbitrary $\epsilon > 0$

$Pr\big(\sup_{\|v\|_2=C} \mathbf{Q}(\hat{\beta}, \theta_0 + v\eta_n) - \mathbf{Q}(\hat{\beta}, \theta_0)\ \big) > 1 - \epsilon$ here $\eta = O_p(\sqrt{\frac{1}{\sum_1^T P_t n}})$

$$\mathbf{Q}(\hat{\beta}, \theta_0 + v\eta_n) - \mathbf{Q}(\hat{\beta}, \theta_0) = \underbrace{\mathbf{Q}(\beta_0, \theta_0 + v\eta_n) - \mathbf{Q}(\beta_0, \theta_0)}_{1}$$
$$+ \underbrace{\mathbf{Q}(\hat{\beta}, \theta_0 + v\eta_n) - \mathbf{Q}(\beta_0, \theta_0 + v\eta_n) - \Big(\mathbf{Q}(\hat{\beta}, \theta_0) - \mathbf{Q}(\beta_0, \theta_0)\Big)}_{2}$$

Expanding term (1) by Taylor series we obtain

$$\underbrace{\eta_n \sum_{r=1}^{q} \frac{\partial \mathbf{Q}(\beta_0, \theta_0)}{\partial \theta_r} v_r}_{11} + \underbrace{\eta_n^2 \sum_{r,r'=1}^{q,q} v_{r'}[-\tilde{t}_{r,r'}(\theta_0)]v_r}_{12} + \underbrace{\eta_n^2 \sum_{r,r'=1}^{q,q} v_{r'}\left[\frac{\partial^2 \mathbf{Q}(\beta_0, \theta^*)}{\partial \theta_r' \theta_r} + \tilde{t}_{r,r'}(\theta_0)\right]v_r}_{13}$$

$\frac{\partial \mathbf{Q}(\beta_0,\theta_0)}{\partial \theta_r} = -(\boldsymbol{Z} - \tilde{X}\boldsymbol{\beta}_0)^T\tilde{\Sigma}^r(\theta_0)^{-1}(\boldsymbol{Z} - \tilde{X}\boldsymbol{\beta}_0) + \text{trace}(\tilde{\Sigma}(\theta_0)\tilde{\Sigma}^r(\theta_0)^{-1})$ Using Markov inequality

$= O_p\left(\sqrt{trace\, \tilde{\Sigma}(\theta_0)\tilde{\Sigma}^r(\theta_0)^{-1}\tilde{\Sigma}(\theta_0)\tilde{\Sigma}^r(\theta_0)^{-1}}\right) = O_p(\|\tilde{\Sigma}(\theta_0)\tilde{\Sigma}^r(\theta_0)^{-1}\|_F)$

$\le O_p(\sqrt{nPT})\lambda_{max}(\Sigma(\theta_0))\lambda_{max}(\Sigma^r(\theta_0)^{-1}) = O_p(\sqrt{nPT})$

$(11) = O_p(1)\|\nu\|$ Let $\tilde{t}_{i,j}(\theta) = trace\big(\mathbb{I}_n \otimes \underbrace{\Sigma^{-1}(\theta_0)\Sigma_i(\theta_0)\Sigma^{-1}(\theta_0)\Sigma_j(\theta_0)}_{t_{i,j}(\theta_0)}\big)$

$= n\, t_{i,j}(\theta_0) = n\, t_{i,i}^{\frac{1}{2}}\left(\frac{t_{i,j}}{\sqrt{t_{i,i}t_{j,j}}}\right)t_{j,j}^{\frac{1}{2}}$

$= n\, t_{i,i}^{\frac{1}{2}} a_{i,j} t_{j,j}^{\frac{1}{2}}$

By assumption, $\lim_n A$ exists with $\lim_n \lambda_{max}(A) = O_p(1)$ and $t_{i,i}(\theta_0)$

$\ge PT\lambda^2_{min}(\Sigma^{-1}(\theta_0))\lambda^2_{min}(\Sigma_i(\theta_0))$

Hence for some constant $K_n$

$$(12) \leq -nPT\eta_n^2\|v\|_2^2 \lim_n \lambda_{max}(A)\lambda_{min}^2(\Sigma^{-1}(\theta_0)) \max_{i\leq q} \lambda_{min}^2(\Sigma_i(\theta_0)) = -nPT\eta_n^2\|\nu\|_2^2 K_n$$

$$\frac{\partial^2 \mathbf{Q}(\beta_0,\theta^*)}{\partial\theta_r'\theta_r} = \underbrace{\text{trace}(\tilde{\Sigma}(\theta^*)\tilde{\Sigma}^{r,r'}(\theta^*)^{-1}) \; - (\boldsymbol{Z}-\tilde{X}\boldsymbol{\beta})^T\tilde{\Sigma}^{r,r'}(\theta^*)^{-1}(\boldsymbol{Z}-\tilde{X}\boldsymbol{\beta})}_{131}$$

$$+ \text{trace}(\tilde{\Sigma}_r(\theta^*)\tilde{\Sigma}^{r'}(\theta^*)^{-1})$$

Similar to expression (11)

$$(131) = O_p(\|\tilde{\Sigma}(\theta^*)\tilde{\Sigma}^{r,r'}(\theta^*)^{-1}\|_F) + \text{trace}\tilde{\Sigma}^{r,r'}(\theta^*)^{-1}\left(\tilde{\Sigma}(\theta^*) - \tilde{\Sigma}(\theta_0)\right)$$

$$\leq O_p(\sqrt{nPT})\lambda_{max}(\Sigma^{r,r'}(\theta^*)^{-1})\lambda_{max}(\Sigma(\theta_0))$$

$$+ n\lambda_{max}[(\Sigma^{r,r'}(\theta^*)^{-1})\, trace(\Sigma(\theta_0) - \Sigma(\theta^*)\,)]$$

$$= O_p(\sqrt{nPT}) + nO_p(1)PT \max_{i\leq PT}\langle \tfrac{\partial\gamma_i}{\partial\theta}\big|_{\theta^{**}}, \theta_0 - \theta^*\rangle$$

$$\leq O_p(\sqrt{nPT}) + nO_p(1)PT \max_{i\leq PT}\|\tfrac{\partial\gamma_i}{\partial\theta}\big|_{\theta^{**}}\|_\infty\|\theta_0 - \theta^*\|_2 = O_p(\sqrt{nPT})$$

Using identity $\tilde{\Sigma}_r(\theta)\tilde{\Sigma}^{r'}(\theta)^{-1} = -\tilde{\Sigma}_r(\theta)\tilde{\Sigma}^{-1}(\theta)\tilde{\Sigma}_{r'}(\theta)\tilde{\Sigma}^{-1}(\theta)$ we obtain that

$$trace(\tilde{\Sigma}_r(\theta_0)\tilde{\Sigma}^{r'}(\theta_0)^{-1}) - trace(\tilde{\Sigma}_r(\theta^*)\tilde{\Sigma}^{r'}(\theta^*)^{-1})$$

$$= trace(\tilde{\Sigma}_r(\theta_0)(\tilde{\Sigma}^{r'}(\theta_0)^{-1} - \tilde{\Sigma}^{r'}(\theta^*)^{-1}\,) + trace(\tilde{\Sigma}^{r'}(\theta^*)^{-1}(\tilde{\Sigma}_r(\theta_0) - \tilde{\Sigma}_r(\theta^*)\,)$$

$$\leq n\lambda_{max}(\Sigma^r(\theta_0))trace(\Sigma_{r'}(\theta_0)^{-1} - \Sigma_{r'}(\theta^*)^{-1}\,)$$

$$+ n\lambda_{max}(\Sigma_{r'}(\theta^*)^{-1})\, trace(\Sigma^r(\theta_0) - \Sigma^r(\theta^*)\,)$$

$$= nO_p(1)PT \max_{i\leq PT}\langle \tfrac{\partial\gamma_i^{(r)}}{\partial\theta}\Big|_{\theta^{**}}, \theta_0 - \theta^*\rangle + nO_p(1)PT \max_{i\leq PT}\langle \tfrac{\partial\gamma_{(r',-1)\,i}}{\partial\theta}\Big|_{\theta^{**}}, \theta_0 - \theta^*\rangle$$

$(13) = O_p(\sqrt{nPT})\eta_n^2\|\nu\|_2^2 = O_p(\frac{1}{\sqrt{nPT}})\|2\|_2^2$ Similarly as expression second part of (131). Thus term (1) is dominated by (12)

As for expression (2) it is easy to check that penalty term vanishes and we are left with

$$(\hat{\beta}-\beta_0)^T\tilde{X}^T[\tilde{\Sigma}^{-1}(\theta_0+\nu\eta_n)-\tilde{\Sigma}^{-1}(\theta_0)](\boldsymbol{Z}-\tilde{X}\boldsymbol{\beta}_0) + (\hat{\beta}-\beta_0)^T\tilde{X}^T[\tilde{\Sigma}^{-1}(\theta_0+\nu\eta_n)-\tilde{\Sigma}^{-1}(\theta_0)]\tilde{X}(\hat{\beta}-\beta_0)$$

$$= \eta_n(\hat{\beta}-\beta_0)^T\tilde{X}^T\textstyle\sum_{r=1}^q \tilde{\Sigma}_r(\theta^*)\nu_r(\boldsymbol{Z}-\tilde{X}\boldsymbol{\beta}) + \; (\hat{\beta}-\beta_0)^T\tilde{X}^T\sum_{r=1}^q \tilde{\Sigma}_r(\theta^*)\nu_r\tilde{X}(\hat{\beta}-\beta_0)$$

$= O_p(\|(\hat{\beta}-\beta_0)\|_2\|(\hat{\theta}-\theta_0)\|_2)\sqrt{n\ trace(\Sigma^{-1}(\theta_0)\Sigma^r(\theta_0)^{-1})}+$

$O_p(n\|(\hat{\beta}-\beta_0)\|_2^2\|(\hat{\theta}-\theta_0)\|_2\lambda_{max}\Sigma^r(\theta^*)^{-1}$ Hence (11) dominates all of the rest of the terms, thus appropriate large value on $\|\nu\|$ we prove the statement. ■

**Theorem 2.6.10.** $\min_{1\le j\le q_n}\hat{\beta}_j^{orc}|_{\hat{\theta}} = O(n^{2c})$

*Proof.* Follows $\min_{1\le j\le q_n}\hat{\beta}_j^{orc}|_{\hat{\theta}} \le \min_{1\le j\le q_n}\beta|_{\theta} + O_p(\|\hat{\beta}_j^{orc}|_{\hat{\theta}}-\beta|_{\theta}\|_2)$. By assumption and the result that $O_p(\|\hat{\beta}_j^{orc}|_{\hat{\theta}}-\beta|_{\theta}\|_2) = o_P(n^{2c})$ ■

## 2.7 Algorithm and methodology to obtain optimal solutions

Based on the objective function (2.23), in order to estimate $\hat{\beta}$ and $\hat{\theta}$, we obtain the solutions of derivative of the penalized likelihoods with respect to each of the unknown parameters and iteratively solve until convergence is obtained. Similar to estimating solutions in Zou & Li (2008), as a first step initialization we replace $\Sigma(\hat{\theta})$ with an identity matrix $\mathbf{I}_{pT}$ and obtain the solution for $\beta$ denoted by $\hat{\beta}^{(ini)}$. Given $\hat{\beta}^{(ini)}$, we obtain a solution for $\hat{\theta}^{(0)}$, which is then used to obtain the solution $\hat{\beta^{(0)}}$, which then used to obtain a solution for $\hat{\theta^{(1)}}$. Finally we obtain $\hat{\beta^{(1)}}$ given the estimates $\hat{\theta}^{(1)}$. Similar to already established MLE based methods under the SCAD penalty it has been show with these one-step estimates consistency can be achieved. There is no apparent need to continue to with further iterations and attain convergence in the estimates. Along similar lines theoretical properties of consistency are established for estimates under the assumption that remainder of the parameters are considered to be fixed at any given iteration, based on the solution obtained from the previous step.

Step (1): Let $\Sigma = \mathbb{I}_{pT}$ and solve $\hat{\beta_t}^{(ini)} = \arg\max_{\beta_t}\mathcal{L}(\beta_t, \Sigma = \mathbb{I}_{pT}; Z)$

Step (2): Let $\beta_t = \hat{\beta_t}^{(ini)}$, Solve $\hat{\theta^{(0)}} = \arg\max_{\theta}\mathcal{L}(\hat{\beta_t}^{(ini)}, \theta; Z)$

Step (3): Solve $\hat{\beta_t}^{(0)} = \arg\max_{\beta_t} \mathcal{L}(\beta, \theta = \hat{\theta}^{(0)}; Z)$

Step (4): Solve $\hat{\theta}^{(1)} = \arg\max_{\theta} \mathcal{L}(\beta_t = \hat{\beta_t}^{(0)}, \theta; Z)$

Step (5): Solve $\hat{\beta_t}^{(1)} = \arg\max_{\beta_t} \mathcal{L}(\beta, \theta = \hat{\theta}^{(1)}; Z)$

The final estimates are therefore given by $\hat{\theta}^{(1)}$ and $\hat{\beta_t}^{(1)}$ for the spatio-temporal covariance abd difference in mean parameters respectively.

### 2.7.1 Asymptotic properties of one-step estimates

In order to show properties of consistency for all estimators, we show that consistency is obtained under the assumption that previous estimates of the fixed parameter used to solve the next iteration are also consistent. Based on these notions, consider the following theorem.

**Theorem 2.7.1.** Assuming conditions A1-A10 hold, then for each fixed $t = \{1, .., T\}$,

(i) (Consistency) there exists an estimate $\hat{\beta}_t = (\hat{\beta}_{1,t}, \hat{\beta}_{2,t})^T$ such that, $\|\hat{\beta}_t - \beta_{0,t}\| = O_p(\sqrt{\frac{s_t}{n}})$

(ii) (Sparsity)further, if (i) holds then, $\mathcal{Q}((\hat{\beta}_1, 0)', \theta; \mathbf{Z}) > argmax_{\|\hat{\beta}_2\|_2 = c\sqrt{\frac{s}{n}}} \mathcal{Q}((\hat{\beta}_1, \hat{\beta}_2)', \theta; \mathbf{Z})$

*Proof.*Check appendix.

In a similar way, we are able to establish existence and consistency results for $\hat{\theta}$ as well. That is,

**Theorem 2.7.2.** Assuming 2.5to (A10) hold, $\|\hat{\theta} - \theta_0\| = O_p(\sqrt{\frac{1}{pTn}})$ as $p, T, n \to \infty$.

Since for every time component we are able to show asymptotic properties, these results can be neatly combined to obtain the following result.

**Theorem 2.7.3.** Using the above theorem one can easily check that for

$\beta_{0,pT\times 1} = (\beta_{0,1,p\times 1}, \beta_{0,2,p\times 1}, ...\beta_{0,t,p\times 1}, ..\beta_{0,T,p\times 1})^T$ and

$\hat{\beta}_{0,pT\times 1} = (\hat{\beta}_{0,1,p\times 1}, \hat{\beta}_{0,2,p\times 1}, ...\hat{\beta}_{0,t,p\times 1}, ..\hat{\beta}_{0,T,p\times 1})^T$

(i) (Consistency) there exists an estimate $\hat{\beta}$ such that $\|\hat{\beta} - \beta_0\| = O_p(\sqrt{\frac{s}{n}})$ as $n \to \infty$.

(ii) (Sparsity) $P(\hat{\beta}_{t,2} = 0 \ \forall t) \to 1$ as $n \to \infty$.

All the proofs of theorems 2.7.1, 2.7.2 and 2.7.3 are provided in the Appendix **??**. The covariance function is assumed to be stationary and the results included are based on the SCAD penalty resulting in unbiased estimators.

## 2.8 Computational Complexity

Thus far, we have described a penalized LDA technique, that can simultaneously estimate parameters of an underlying spatio-temporal covariance model and selection and estimation of the differences between the means. However, whenever the vector $Z$ is constructed the corresponding matrix $\tilde{\Sigma}(\theta)$ which is of dimension $(n-1)pT \times (n-1)pT$. Assuming the subjects are independent we will require to calculate the inverse of a $pT \times pT$ matrix, which could have a computational cost of $O(pT^3)$. In order to ease this burden, we begin by studying tapering techniques.

### 2.8.1 Covariance Tapering

In spatial statistics Kaufman et al. (2008) introduced covariance tapering to approximate the likelihood by replacing the spatial covariance with a positive-definite tapered version of the covariance matrix and established the computational gain while maintaining the the underlying theoretical properties. For very large datasets, Furrer et al. (2006) delineate the use of tapered covariance matrices. As the density of sites in a image is exponentially larger than the number of time-points in our longitudinal imaging study we taper only the spatial covariance of a separable spatio-temporal covariance matrix.

Suppose we use tapering function $K_{Tap}(h, w)$ only for spatial domain with the spectral density of

tapered function

$$f_K(\omega) = (2\pi)^{-d} \int_{\mathbb{R}^d} \exp(-iw^T x) K_{Tap}(x, \omega) dx$$

with the restriction, $f_K(\omega) \leq \frac{M_\psi}{(1+\frac{\|\omega\|^2}{\alpha^2})^{\nu+d/2+\psi}}$ where $\psi > (1-\nu, d/4)$. This tapering function also included Wedland taper function of degree $k$ satisfies the above condition 2.8.1 for $k > max\{1/2, \nu + (d-2)/4 + \delta\}$ for some $\delta > 0$. However, we consider the Wedland taper function for each time fixed covariance matrix $\Sigma_t$ for $K_{Tap}(h, r_t) = [(1 - \frac{h}{r_t})_+]^2$ have $k = 2, d = 3$ here

### 2.8.2 One way Tapering vs Two way Tapering

One way tapering yields biased score function:

$l_1(\theta) = -\frac{nPT}{2} \log(2\pi) - (Z - \tilde{X}\beta)^T (\tilde{\Sigma} \circ K_{Tap}^{-1})(Z - \tilde{X}\beta) - \log(\det(\tilde{\Sigma} \circ K_{Tap}))$

$- \sum_{t=1}^{T} \sum_{j=1}^{P} p_{\lambda_{t,n}}(|\beta_{0\,t,j}|)$

Two way tapering yields unbiased score function:

$l_2(\theta) = -\frac{nPT}{2} \log(2\pi) - (Z - \tilde{X}\beta)^T [(\tilde{\Sigma} \circ K_{Tap}^{-1}) \circ K_{Tap}](Z - \tilde{X}\beta) - \log(\det(\tilde{\Sigma} \circ K_{Tap}))$

$- \sum_{t=1}^{T} \sum_{j=1}^{P} p_{\lambda_{t,n}}(|\beta_{0\,t,j}|)$

It is established that both methods give asymptotically close estimates $\|\hat{\theta}_{1,taper} - \hat{\theta}_{2,taper}\|_2 = o_p(\sqrt{\frac{1}{\sum_{t=1}^{T} P_t n}})$ under certain assumptions.

### 2.8.3 Tapering range

We have established that, choosing tapering range guarantees estimation consistency of $\theta_0$ as $n \to \infty$ under increasing domain setup. In the work of Furrer et al. (2016) and Chu et al. (2011), authors have established estimation consistency under some sufficient conditions;which are satisfied for our Matern class of covariance. Mainly due to the fact that for appropriate true parameters $\gamma(s, t; \theta) < \frac{A}{1+\|s\|^{3+\alpha_1}+|t|^{1+\alpha_2}}$ and its derivative $\frac{\partial \gamma(s,t;\theta)}{\partial \theta} \frac{A}{1+\|s\|^{3+\alpha_1}+|t|^{1+\alpha_2}}$ for some $\alpha_1 > 0$ and $\alpha_2 > 0$.

Similar results hold for non separable covariances as well. Additionally, $\int_0^\infty s^{d+1}t^2\gamma(s,t)dsdt < \infty$ as spectral density is differentiable indefinitely.Thus we show that tapering range $r_t = O(\sqrt{P_t})$

$$\sup_{\theta_0\in\Theta} |\mathbf{L}_{\theta_0} - \mathbf{L}_{\theta_{tap}}| = \frac{1}{n}\log(|\tilde{\Sigma}(\theta_0)(\tilde{\Sigma}(\theta_0)\circ K_{taper})^{-1})| + \frac{1}{n}(\boldsymbol{Z}-X\beta_0)^T(\tilde{\Sigma}^{-1}(\theta_0) - \tilde{\Sigma}(\theta_0)\circ K_{taper})^{-1})(\boldsymbol{Z}-X\beta_0) = o_p(1)$$

## 2.9 Misclassification Optimality

This theorem shows that after penalized maximum likelihood estimation under tapering, the plug in classifier is optimal or universal.

**Theorem 2.9.1.** Suppose $\frac{s_t}{nC_p} \to 0$ the worst classification error rate of

$W(\hat{\delta}) \to 1 - \Phi(\frac{\sqrt{C_0}}{2})$

*Proof.* Refer to appendix 2.11.4.1 ■

## 2.10 MRI Data Preprocessing

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). For up-to-date information, see www.adni-info.org. To demonstrate the methodology we consider MRI images obtained from a two year study using a 3T scanner. A total of 51 subjects (31 Controls, 18 AD ) are collected. On average each of these subjects have 4 visits. The preprocessing protocol from ADNI allows us to download N3 (nonparametric nonuniformity normalization) bias-corrected NIFTi images. For visualizing these images, we use the freely available toolkit ITKsnap developed by Yushkevich et al. (2016). An affine registration with the skull on was performed on every subject to its separate visits using ANTs Avants et al. (2011) registration implemented in R using the AntsR and extrantsr packages Muschelli et al. (2018). i.e. visit two, three and four were affine registered to visit one.

After the first set of registrations on this longitudinal dataset, for each of the subjects using just the first visit, skull-stripping was performed by extracting only the brain tissue using the tool MASS developed by Doshi et al. (2013). This was run in parallel as it is known to be computationally heavy. If this is hard to implement one may consider the simpler FSL Bet tool developed by Popescu et al. (2012).

The brain mask obtained after skull stripping the first visit corresponding to each subject, is then applied to all other visits of the respective subject. To perform voxel-wise analyses, a deformable registration is performed using a template. For the purposes of the 3D 3T MRI scans we chose the SRI24 atlas Rohlfing et al. (2008). This deformable registration was once again performed using ANTs registration. These transformations were calculated using the first visit of each subject to the template. A forward transformation consists of .mat file (for initial affine registration) and a forward warp image .nii.gz file. The transformations then may be applied to the corresponding time points. A quality control check may then be done using the ITKsnap tool to ensure that all images have been registered to the same space.

In order to obtain ROI labels one can visit (www.nitrc.org/projects/sri24) to obtain the segmentation of the ROIs and the corresponding label text file. The hippocampus for AD studies is labeled as 37 (left) and 38(right).

## 2.11 Simulation Studies

In the following simulation study, we investigate a variety of conditions under which the proposed methods are employed. On a lattice grid of $p \times p \in \mathbb{R}^2$ multinomial gaussian observations where $p = \{10, 15, 20\}$ are produced over $t = \{1, .., 4\}$ time points. The mean of this multinomial gaussian distribution is characterized in two ways. The first condition assumes that there is a strong signal indicating a significant difference between the two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ i.e.

$\Delta = \{1, ..., 1, 2....2, 0..., 0\}^T$ where $s = 20$ is the number of non-zero entities and the first 10 are 1 while the following 10 are 2. Similarly a smaller difference between the two classes was assumed with $\Delta_{small} = \{0.75, ..., 0.75, 0.5....0.5, 0..., 0\}^T$. Under space and time separable assumption multiple spatial covariance functions are assumed such as:

- **Exponential covariance**: The spatial dependence of the error terms are $\gamma(d) = \sigma^2(1 - c)\exp(-d/r)$ where $\gamma(d) = \sigma^2$ when $d = 0$ and $d$ denotes the euclidean distance between two spatial points. $r = \{3, 10\}$, $c = \{0.2, 0.5\}$ and $\sigma = 1$.

- **Polynomial correlation**: The covariance function is $\gamma(d) = \rho^d$ where $\rho = 0.9$.

- **Matern covariance**: The covariance function is given by, $\gamma(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)}\big(\sqrt{2\nu\frac{d}{\rho}}\big)^{\nu}\mathcal{K}_{\nu}\big(\sqrt{2\nu\frac{d}{\rho}}\big)$ where $\mathcal{K}_{\nu}(\cdot)$ is the modified Bessel function of the second kind, $\rho, \nu > 0$ and $\Gamma$ is the gamma function. In this study, $\nu = 2$ and $\rho = 5$.

The covariance assumed for the four time points was exponential where $\gamma(d_t) = \exp(-d_t/rt)$ where $d_t$ denotes the euclidean distance between pairs of time points and $rt = \{0.5, 0.7, 1\}$.
For the non-separable case we use the following covariance function,

- **Gneiting covariance**: This non-separable space time covariance is, $\gamma(d, d_t) = \frac{e^{\left(\frac{-d^{\nu}}{(1+d_t^{\lambda})^{0.5\gamma\mu}}\right)}}{1+d_t^{\lambda}}$, where $\nu, \lambda \in [0, 2]$ and $\gamma \in (0, 1)$. If $\gamma = 0$ then the model is separable. In the simulation study the values considered were $\nu = 0.8$, $\lambda = 1$ and $\gamma = 0.6$.

For all of the simulation studies we have 100, 225 and 400 locations on 4 time points. Therefore the total number of observations per subject is 400, 900 and 1600 observations respectively. The total number of non-zero components in the mean vector $s = 20$. So best case scenario for selection is when true positives (TP) is 20 and false positives (FP) is 0. A total of 100 subjects per study is generated with a 50% training and validation split.

### 2.11.1 Exponential Space Time Covariance (weak correlations)

#### 2.11.1.1 With $\Delta$

| $p \times p$ | r=3 | c=0.2 | $\sigma = 1$ | $rt = 1$ |
|---|---|---|---|---|
| 100 | 3.009 (0.0979339536) | 0.231 (0.0228625958) | 1 (4.234e-07) | 0.992 (0.0103475934) |
| 225 | 2.996 (0.0736011373) | 0.234 (0.0175467671) | 1 (5.015e-07) | 0.995 (0.0069044814) |
| 400 | 3.002 (0.0516560637) | 0.233 (0.0125129651) | 1 (2.954e-07) | 0.996 (0.0048776816) |

Table 2.1: Exponential separable space and time covariance estimation when $\Delta$ is the mean vector

| $p \times p$ | TP | FP | MSE | Train1 | Train2 | Test1 | Test2 |
|---|---|---|---|---|---|---|---|
| 100 | 20.00 | 47.14 | 0.00202 | 1.00 | 1.00 | 1.00 | 1.00 |
| 225 | 20.00 | 111.02 | 0.00124 | 1.00 | 1.00 | 1.00 | 1.00 |
| 400 | 19.99 | 189.68 | 0.00097 | 1.00 | 1.00 | 1.00 | 1.00 |

Table 2.2: Estimation and selection of $\Delta$ using the estimated covariance

| $p \times p$ | TP | FP | MSE | Train1 | Train2 | Test1 | Test2 |
|---|---|---|---|---|---|---|---|
| 100 | 19.87 | 45.38 | 0.00851 | 1.00 | 1.00 | 1.00 | 1.00 |
| 225 | 19.83 | 64.56 | 0.00467 | 1.00 | 1.00 | 1.00 | 1.00 |
| 400 | 19.70 | 78.59 | 0.00310 | 1.00 | 1.00 | 1.00 | 1.00 |

Table 2.3: Estimation and selection of $\Delta$ under independence

| $p \times p$ | r=3 | c=0.2 | $\sigma = 1$ | $rt = 1$ |
|---|---|---|---|---|
| 100 | 2.992 (0.0940290558) | 0.234 (0.0215993179) | 1 (5e-07) | 0.993 (0.0109421777) |
| 225 | 3.005 (0.0735975377) | 0.232 (0.0166009694) | 1 (4.172e-07) | 0.993 (0.0080306034) |
| 400 | 3 (0.0567520396) | 0.233 (0.0131230581) | 1 (4.156e-07) | 0.995 (0.0060978283) |

Table 2.4: Exponential separable space and time covariance estimation when $\Delta_{small}$ is the mean vector

| $p \times p$ | TP | FP | MSE | Train1 | Train2 | Test1 | Test2 |
|---|---|---|---|---|---|---|---|
| 100 | 17.66 | 119.99 | 0.00839 | 0.79 | 0.79 | 0.79 | 0.79 |
| 225 | 16.13 | 231.17 | 0.00506 | 0.73 | 0.73 | 0.73 | 0.73 |
| 400 | 12.99 | 311.30 | 0.00385 | 0.67 | 0.67 | 0.67 | 0.66 |

Table 2.5: Estimation and selection of $\Delta_{small}$ using the estimated covariance

| $p \times p$ | TP | FP | MSE | Train1 | Train2 | Test1 | Test2 |
|---|---|---|---|---|---|---|---|
| 100 | 15.70 | 45.50 | 0.01081 | 0.75 | 0.75 | 0.75 | 0.75 |
| 225 | 14.41 | 55.40 | 0.00561 | 1.00 | 1.00 | 0.71 | 0.72 |
| 400 | 12.68 | 54.21 | 0.00345 | 0.69 | 0.69 | 0.69 | 1.00 |

Table 2.6: Estimation and selection of $\Delta_{small}$ under independence

### 2.11.2 Exponential Space Time Covariance (strong correlations)

#### 2.11.2.1 With $\Delta$

| $p \times p$ | r=10 | c=0.5 | $\sigma = 1$ | rt=0.5 |
|---|---|---|---|---|
| 100 | 10.024 (0.6592026826) | 0.519 (0.0310983865) | 1 (3.3601e-06) | 0.47 (0.1196661211) |
| 225 | 9.918 (0.4420207073) | 0.523 (0.021304777) | 1 (2.4351e-06) | 0.466 (0.1284748148) |
| 400 | 9.952 (0.3472843583) | 0.522 (0.0165295756) | 1 (3.5861e-06) | 0.475 (0.1094737503) |

Table 2.7: Exponential separable space and time covariance estimation when $\Delta$ is the mean vector

| $p \times p$ | TP | FP | MSE | Train1 | Train2 | Test1 | Test2 |
|---|---|---|---|---|---|---|---|
| 100 | 20.00 | 0.79 | 0.00029 | 1.00 | 1.00 | 0.79 | 1.00 |
| 225 | 16.13 | 231.17 | 0.00506 | 0.73 | 0.73 | 0.73 | 0.73 |
| 400 | 12.99 | 311.30 | 0.00385 | 0.67 | 0.67 | 0.67 | 0.66 |

Table 2.8: Estimation and selection of $\Delta$ using the estimated covariance

| $p \times p$ | TP | FP | MSE | Train1 | Train2 | Test1 | Test2 |
|---|---|---|---|---|---|---|---|
| 10x10 | 19.98 | 47.08 | 0.00487 | 1.00 | 1.00 | 1.00 | 1.00 |
| 15x15 | 19.98 | 61.14 | 0.00229 | 1.00 | 1.00 | 1.00 | 1.00 |
| 20x20 | 19.99 | 78.91 | 0.00154 | 1.00 | 1.00 | 1.00 | 1.00 |

Table 2.9: Estimation and selection of $\Delta$ under independence

| $p \times p$ | r= 10 | c=0.5 | $\sigma = 1$ | rt=0.5 |
|---|---|---|---|---|
| 100 | 10.056 (0.5445091578) | 0.517 (0.0251724131) | 1 (2.5151e-06) | 0.466 (0.1292439526) |
| 225 | 10.021 (0.415936012) | 0.519 (0.0190434918) | 1 (5.481e-07) | 0.46 (0.136410963) |
| 400 | 9.952 (0.3829308584) | 0.522 (0.0185292498) | 1 (2.2068e-06) | 0.475 (0.1505265743) |

Table 2.10: Exponential separable space and time covariance estimation when $\Delta_{small}$ is the mean vector

| $p \times p$ | TP | FP | MSE | Train1 | Train2 | Test1 | Test2 |
|---|---|---|---|---|---|---|---|
| 100 | 20.00 | 29.41 | 0.00036 | 1.00 | 1.00 | 1.00 | 1.00 |
| 225 | 19.99 | 231.17 | 0.00021 | 1.00 | 1.00 | 1.00 | 1.00 |
| 400 | 19.99 | 102.89 | 0.00015 | 1.00 | 1.00 | 1.00 | 1.00 |

Table 2.11: Estimation and selection of $\Delta_{small}$ using the estimated covariance

| $p \times p$ | TP | FP | MSE | Train1 | Train2 | Test1 | Test2 |
|---|---|---|---|---|---|---|---|
| 100 | 19.98 | 62.76 | 0.00630 | 1.00 | 1.00 | 1.00 | 0.99 |
| 225 | 19.98 | 61.14 | 0.00375 | 0.97 | 0.97 | 0.97 | 0.98 |
| 400 | 19.99 | 78.91 | 0.00225 | 0.98 | 0.98 | 0.98 | 0.98 |

Table 2.12: Estimation and selection of $\Delta_{small}$ under independence

### 2.11.3 Matern Space Covariance and exponential time (separable)

#### 2.11.3.1 With $\Delta$

| $p \times p$ | $\nu = 2$ | rt=0.7 | $\phi = 5$ |
|---|---|---|---|
| 100 | 1.998 (0.0143793046) | 0.698 (0.0175455886) | 4.795 (0.1068378951) |
| 225 | 2.001 (0.0091730859) | 0.701 (0.0108351911) | 4.808 (0.06846612) |
| 400 | 2 (0.0070390667) | 0.7 (0.0074068311) | 4.799 (0.0494941786) |

Table 2.13: Matern covariance with separable exponential time when $\Delta$ is the mean vector

| $p \times p$ | TP | FP | MSE | Train1 | Train2 | Test1 | Test2 |
|---|---|---|---|---|---|---|---|
| 100 | 15.18 | 86.51 | 0.03 | 0.97 | 0.97 | 0.97 | 0.98 |
| 225 | 14.11 | 155.46 | 0.01 | 0.97 | 0.97 | 0.97 | 0.97 |
| 400 | 12.73 | 214.40 | 0.01 | 0.97 | 0.97 | 1.00 | 0.97 |

Table 2.14: Estimation and selection of $\Delta$ using the estimated covariance

| $p \times p$ | TP | FP | MSE | Train1 | Train2 | Test1 | Test2 |
|---|---|---|---|---|---|---|---|
| 100 | 14.49 | 62.76 | 0.06 | 0.93 | 0.93 | 0.93 | 0.93 |
| 225 | 17.32 | 61.14 | 0.00 | 0.97 | 0.97 | 0.97 | 0.98 |
| 400 | 10.19 | 51.06 | 0.02 | 0.83 | 0.83 | 0.98 | 0.84 |

Table 2.15: Estimation and selection of $\Delta$ under independence

### 2.11.4 Non separable space and time Gneiting covariance (separable)

#### 2.11.4.1 With $\Delta$

| $p \times p$ | $\nu = 0.8$ | $\lambda = 1$ | $\gamma = 0.6$ |
|---|---|---|---|
| 100 | 0.831 (0.0147602785) | 1.059 (0.0263549226) | 0.982 (0.0063041298) |
| 225 | 0.832 (0.0108536628) | 1.057 (0.0162489845) | 0.983 (0.0039539343) |
| 400 | 0.831 (0.0078036784) | 1.057 (0.0130021963) | 0.983 (0.0031399539) |

Table 2.16: Gneiting covariance with non-separable when $\Delta$ is the mean vector

| $p \times p$ | TP | FP | MSE | Train1 | Train2 | Test1 | Test2 |
|---|---|---|---|---|---|---|---|
| 100 | 19.08 | 111.03 | 0.03 | 0.97 | 0.97 | 0.97 | 0.96 |
| 225 | 17.22 | 178.08 | 0.01 | 0.96 | 0.96 | 0.96 | 0.96 |
| 400 | 14.40 | 178.88 | 0.01 | 0.95 | 0.95 | 0.95 | 0.95 |

Table 2.17: Estimation and Selection of $\Delta$ using the estimated covariance

# APPENDIX

# APPENDIX

**Useful lemmas**

**Lemma 2.11.1.** If $X \sim N(0,1)$ then, $X = O_p(1)$

*Proof.* Using Gaussian concentration inequality $P(|X| > \frac{\sqrt{-log(\epsilon/2)}}{2}) \leq \epsilon$ for every $\epsilon > 0$ there exists a $\delta = \frac{\sqrt{-log(\epsilon/2)}}{2}$. ■

**Lemma 2.11.2.** If $X \sim \chi^2(K)$ then, $X = O_p(K)$

*Proof.* Using the sub-exponential concentration inequality $X-K$ is subexponential with parameters $(4K, 2)$. Similarly for every $\epsilon > 0$ there exist a $\delta$ which implies $P(\frac{|X-K|}{4K} > \delta_\epsilon) < \epsilon$ for some constant $K > 0$. ■

**Theorem 2.11.3** (2.7.2)**.** Assuming $\|\hat{\theta} - \theta\|_2 = O_p(\sqrt{\frac{1}{PTn}})$ , we want to prove that for each $t = 1, 2..., T$ $\|\hat{\beta}_t - \beta_{t,0}\|_2 = O_p(\sqrt{\frac{s_t}{n}})$ and $\hat{\beta}_{t\,2} = 0_{P\times 1}$ with probability close to 1 where $\hat{\beta} = \arg\max_\beta \mathbf{L}(\beta, \hat{\theta}; Z)$

*Proof.* For each fixed $t$ the proof consists of two steps
In first part we show that $\|\hat{\beta}_t - \beta_{t,0}\|_2 = O_p(\sqrt{\frac{s_t}{n}})$ . In the next part we show that $\hat{\beta}_{t\,2} = 0_{P\times 1}$ with probability close to 1 . To prove this second part we need an additional lemma 2.11.4 from the paper

The proof of first part

Denote vector $u_t \in \mathbb{R}^{p_n}$ with entries 1 corresponding to index j such that $\beta_{t,j} \neq 0$ and 0 elsewhere, but without loss of generality , we denote $u_t$ as $u$.Similarly, $X_t$ denotes the colomns of design matrix $X$ corresponding to $\beta_{t,j} \neq 0$. Hence we get identity that $X_t u_t = X \begin{pmatrix} \mathbb{I}_{s_t \times s_t} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} u$ It is sufficient to show that for any $\epsilon > 0$ ,$\sum_{t=1}^{T} s_t = s$, $\xi_{n,t} = O_p(\sqrt{\frac{s_t}{n}})$ , $\xi_n = O_p(\sqrt{\frac{s}{n}})$ and for some $C > 0$ under the condition that for $r \in \{1, 2, ....T\} - \{t\}$ we have $\|\hat{\beta}_r - \beta_r\|_2 = O_p(\sqrt{\frac{s_r}{n}})$

$$Pr\big( \sup_{\|u\|_2=C} \mathbf{Q}(\beta_{t,0} + u\xi_{n,t}, \hat{\theta}) - \mathbf{Q}(\beta_{t,0}, \hat{\theta}) \ \big) > 1 - \epsilon$$

or equivalently we can prove that

$$Pr\big( \sup_{\|u\|_2=C} \mathbf{Q}(\beta_{0,t} + u\xi_{n}, \hat{\theta}) - \mathbf{Q}(\beta_{0,t}, \hat{\theta}) \ \big) > 1 - \epsilon$$

$$\begin{aligned}
&\mathbf{Q}(\beta_{0,t} + u\xi_n, \hat{\theta}) - \mathbf{Q}(\beta_{0,t}, \hat{\theta}) \\
&= \mathbf{L}(\beta_{0,t} + u\xi_n, \hat{\theta}) - \mathbf{L}(\beta_{0,t}, \hat{\theta}) - n\sum_{j=1}^{P_t}(p_{\lambda_{n,t}}(|\beta_{t,0,j} + u\xi_n|) - p_{\lambda_{n,t}}(|\beta_{t,0,j}|)) \\
&= \mathbf{L}(\beta_{0,t} + u\xi_n, \hat{\theta}) - \mathbf{L}(\beta_{0,t}, \hat{\theta}) - n\sum_{j=1}^{s_t}(p_{\lambda_{n,t}}(|\beta_{t,0,j} + u\xi_n|) - p_{\lambda_{n,t}}(|\beta_{t,0,j}|)) \\
&- n\sum_{j=s_t+1}^{p_t}(p_{\lambda_{n,t}}(|\beta_{t,0,j} + u\xi_n|) - p_{\lambda_{n,t}}(|\beta_{t,0,j}|)) \\
&\leq \mathbf{L}(\beta_{0,t} + u\xi_n, \hat{\theta}) - \mathbf{L}(\beta_{0,t}, \hat{\theta}) - n\sum_{j=1}^{s_t}(p_{\lambda_{n,t}}(|\beta_{t,0,j} + u\xi_n|) - p_{\lambda_{n,t}}(|\beta_{t,0,j}|))
\end{aligned}$$

for$j = s_t + 1, ., p_t$

$p_{\lambda_{n,t}}(|\beta_{t,0,j}| = 0) = 0$

$p_{\lambda_{n,t}}(|\beta_{t,0,j} - u\xi_n| > 0) = 0$

$$\begin{aligned}
&= \frac{1}{2}u_t^T X_t^T \tilde{\Sigma}(\hat{\theta})^{-1} X_t u_t \xi_n -^T \tilde{\Sigma}(\hat{\theta})^{-1} X_t u_t \\
&- n\sum_{j=1}^{s_t}(p_{\lambda_{n,t}}(|\beta_{t,0,j} + u\xi_n|) - p_{\lambda_n}(|\beta_{t,0,j}|)) \\
&= \frac{1}{2}u_t^T X_t^T \tilde{\Sigma}(\hat{\theta})^{-1} X_t u_t \xi_n - (Z - X\beta)^T \tilde{\Sigma}(\hat{\theta})^{-1} X_t u_t \\
&- n\sum_{j=1}^{s_t}(p_{\lambda_n}(|\beta_{t,0,j} + u\xi_n|) - p_{\lambda_n}(|\beta_{t,0,j}|)) \\
&= \underbrace{\frac{1}{2}u_t^T X_t^T \tilde{\Sigma}(\hat{\theta})^{-1} X_t u_t \xi_{n,t}^2}_{1} - \underbrace{(Z - X\beta)^T \tilde{\Sigma}(\hat{\theta})^{-1} X_t u_t \xi_{n,t}}_{2}
\end{aligned}$$

$$\underbrace{-n\sum_{j=1}^{s_t}\left(p'_{\lambda_{n,t}}(|\beta 0,j|)\mathrm{sgn}(\beta 0,j)u_t\xi_{n,t}\right.}_{3a}+\underbrace{\left.p''_{\lambda_{n,t}}(|\beta_{0,j}|)u_t^2\xi_{n,t}^2(1+o_p(1))\right)}_{3b}$$

For term (1) we get using eigenvalue inequality result that $\psi_i$ is $i$ th eigenvalue

*Proof.* $\underbrace{\frac{1}{2}u_t^TX_t^T\tilde{\Sigma}(\hat{\theta})^{-1}X_tu_t\xi_n}_{1}=\underbrace{\frac{1}{2}u^TX^T\tilde{\Sigma}(\theta_0)^{-1}X_tu_t\xi_n}_{1a}$

$+\ \underbrace{\frac{1}{2}u_t^TX_t^T\sum_{r=1}^{q}u^TX^T\dot{\tilde{\Sigma}}(\theta^*)^r(\hat{\theta}_r-\theta_{0,r})u^TX^T\xi_n}_{1b}$

$1(a)\leq O(\|u\|^2X^TX\min_i\psi_i(\tilde{\Sigma}(\hat{\theta})^{-1}))\xi_n^2=O(\min_i\psi_i(\tilde{X}^T(\mathbf{I}_{n-1}+\mathbf{J}_{n-1}))\tilde{X})\min_i\psi_i(\tilde{\Sigma}(\hat{\theta})^{-1}))\xi_n^2$

$=O(n\xi_{n,t}^2)$by eigenvalue inequality

Similarly using CS inequality$1(b)=O(n\xi_{n,t}^2\eta_n)$ ■

For term (2) we use Taylor expansion of $\tilde{\Sigma}(\hat{\theta})^{-1}$ around $\theta_0$ such that $\hat{\theta} - \theta_0 = v\eta_n$

*Proof.* $-(Z - X\beta)^T\tilde{\Sigma}(\hat{\theta})^{-1}X_t u_t = \underbrace{-(Z - X\beta)^T\tilde{\Sigma}(\theta_0)^{-1}Xu\xi_n}_{2a}$

$\underbrace{-\sum_{r=1}^{q}(Z_t - X\beta_t)^T\dot{\tilde{\Sigma}}(\theta^*)^r(\hat{\theta}_r - \theta_{0,r})Xu\xi_n}_{2b}$

term (2a) using Markov inequality that for any r.v

$Z = O_p\sqrt{\mathbb{E}(Z^2)}$ $-(Z - X\beta)^T\tilde{\Sigma}(\theta_0)^{-1}X_t u_t\xi_n = O_p(\sqrt{\min_i \psi_i(\tilde{\Sigma}(\theta_0)^{-1})\, trace(\tilde{X}^T\tilde{X})}\|u\|_2\xi_n)$

$= O_p(\sqrt{\min_i \psi_i(\Sigma(\theta_0)^{-1})\, trace(X_t^T X_t)\begin{pmatrix} -\tau_2\mathbf{1}_{n_1} \\ \tau_1\mathbf{1}_{n_2-1}\end{pmatrix}^T(\mathbf{I}_{n-1} + \mathbf{J}_{n-1})\begin{pmatrix} -\tau_2\mathbf{1}_{n_1} \\ \tau_1\mathbf{1}_{n_2-1}\end{pmatrix}}\|u\|_2\xi_n)$

$= O_p(\sqrt{\frac{n_1 n_2}{n}s_t}\|u\|\xi_n) = O_p(\sqrt{ns_t}\|u\|\xi_n)$ ■

Similarly term (2b) $\sum_{r=1}^{q}(Z - X\beta)^T\tilde{\Sigma}(\theta^*)^r Xu\xi_n$

$= O_p(\sqrt{\min_i \psi_i(\tilde{\Sigma}(\theta^*)^r)\min_i(\tilde{X}^T\tilde{X})}\|u\|_2\|v\|_2\xi_n\eta_n) = O_p(q\sqrt{n_1 n_2}n)$

For term (3a) using CS inequality and using assumption

$-n\sum_{j=1}^{s}(p'_{\lambda_n}(|\beta 0, j|)\text{sgn}(\beta 0, j)u\xi_n \leq n\sqrt{s}\max_j p'_{\lambda_n}(|\beta 0, j|)\|u\|_2^2\xi_n = O(n\sqrt{s}\xi_n\, a_n) = O_p(\sqrt{sn}\xi_n)$

Term (3b)using CS inequality and using assumption $p''_{\lambda_n}(|\beta_{0,j}|)u^2\xi_n^2 \leq n\sqrt{(s)}\xi_n^2 b_n = O_p(n\xi_n^2)$ ■

Here we observe that term 1 dominates all other term hence for appropriate constant $C$ the value above is negative, hence proved.

**Lemma 2.11.4.** This lemma proves sparsity of the estimator.SCAD penalized estimator demonstrates this oracle property which means as if number of zeros in parameter in known initially.Now, let $\hat{\beta}_t = (\hat{\beta}_t^1, \hat{\beta}_t^2)^T$ .For any given $\hat{\beta}_t$ satisfying $\|\hat{\beta}_t^1 - \beta_{t,0}^1\|_2 = O_p(\sqrt{\frac{s_t}{n}})$ and assumptions A(1) to A(12) then with high probability

$$\mathbf{Q}\begin{pmatrix}\hat{\beta}_t^1 \\ \mathbf{0}\end{pmatrix} = \max_{\|\hat{\beta}_t^2\| \leq C\sqrt{\frac{s_t}{n}}}\mathbf{Q}\begin{pmatrix}\hat{\beta}_t^1 \\ \hat{\beta}_t^2\end{pmatrix}$$

*Proof.* It is sufficient to prove that for $j = s_t, s_t + 1, ....p_t$

$$\frac{\partial Q(\beta_t)}{\partial \beta_{t,j}} < 0 \text{ for } 0 < \beta_{t,j} < C\sqrt{\frac{s_t}{n}}$$

and

$$\frac{\partial Q(\beta_t)}{\partial \beta_{t,j}} > 0 \text{ for } 0 > \beta_{t,j} > -C\sqrt{\frac{s_t}{n}}$$

$$\frac{\partial Q(\beta_t;\hat{\theta})}{\partial \beta_{t,j}} = -(Z - X\beta)^T\tilde{\Sigma}(\hat{\theta})^{-1}X_j \ - nP'_\lambda(|\beta_{t,j}|)sgn(\beta_{t,j})$$

$$-(Z - X\beta)^T\tilde{\Sigma}(\hat{\theta})^{-1}X_j = -(Z - X\beta)^T\tilde{\Sigma}(\theta)^{-1}X_j - \sum_{k=1}^{Q}(Z - X\beta)^T\tilde{\Sigma}^k(\theta^*)^{-1}u_k\eta_n$$

$$-(Z - X\beta)^T\tilde{\Sigma}(\theta)^{-1}X_j = O_p(\sqrt{X_j\tilde{\Sigma}^{-1}X_j}) = O_p(\sqrt{n}) \text{ by concentration inequality}$$

similarly,

$$\sum_{k=1}^{Q}(Z - X\beta)^T\tilde{\Sigma}^k(\theta^*)^{-1}u_k\eta_n = O_p(\sqrt{n}\eta_n)$$

Collecting all terms we achieve that

$$\frac{\partial Q(\beta_t;\hat{\theta})}{\partial \beta_{t,j}} = n\lambda_{n,t}\big(O_p(\frac{\sqrt{s_t}}{\sqrt{n}}\lambda_{n,t}) + \frac{p'_{\lambda_{n,t}}(|\beta_j|)}{\lambda_{n,t}}sgn(\beta_j)\big)$$

Since $\frac{\sqrt{s_t}}{\sqrt{n}}\lambda_{n,t} \to 0$ $sgn(\beta_j)$ dominates , hence proved ■

**Lemma 2.11.5.** $\|\hat{\beta} - \beta_0\| = O_p(\sqrt{\sum^T \frac{s_t}{n}})$ then $\|\hat{\theta} - \theta_0\| = O_p(\sqrt{\frac{1}{\sum_1^T P_t n}})$

*Proof.* We will show that for arbitrary $\epsilon > 0$

$Pr\big(\sup_{\|v\|_2=C} \mathbf{Q}(\hat{\beta}, \theta_0 + v\eta_n) - \mathbf{Q}(\hat{\beta}, \theta_0)\ \big) > 1 - \epsilon$ here $\eta = O_p(\sqrt{\frac{1}{\sum_1^T P_t n}})$

By Taylor Series expansion around $\hat{\theta}$

$$\mathbf{Q}(\hat{\beta}, \theta_0 + v\eta_n) - \mathbf{Q}(\hat{\beta}, \theta_0) = \underbrace{\sum_{r=1}^{q} (Z - X\hat{\beta})^T \dot{\tilde{\Sigma}}(\theta_0)^r (Z - X\hat{\beta}) v\eta_2}_{1} +$$

$$\underbrace{\sum_{r,r'=1}^{q} v^T (Z - X\hat{\beta})^T [\ddot{\tilde{\Sigma}}(\theta^*)^{r,r'} - \mathbf{D}](Z - X\hat{\beta}) v\eta_n^2}_{2} +$$

$$\underbrace{v^T (Z - X\hat{\beta})^T \mathbf{D} (Z - X\hat{\beta}) v\eta_n^2}_{3}$$

■

**Lemma 2.11.6.** *$lim_{p \wedge T \to \infty} \lambda_{max} \Sigma(\theta^*) = O(1)$*

*Proof.*

$$\begin{aligned}
&lim_{P,T \to \infty, \infty} \lambda_{max} \Sigma(\theta^*) \\
&\leq \max_i \sum_{j=1}^{PT} |\Sigma_{i,j}| \\
&\leq \sum_{j=1}^{\infty} |\Sigma_{i,j}| \\
&\leq \int_{s=0}^{\infty} \int_{t=0}^{\infty} |\gamma(s,t;\theta)| \\
&\leq \int_{\omega=0}^{\infty} \int_{\tau=0}^{\infty} | \int_{\mathbb{R}^d} \int_{\mathbb{R}} exp(\omega' s + \tau' t) f(\omega, \tau)| \\
&\leq 2 \int_{\|\omega\|=0}^{\infty} \int_{\tau=0}^{\infty} \eta(\alpha^2 \beta^2 + \beta^2 \|\omega\|^2 + \alpha^2 \tau^2 + \epsilon \|\omega\|^2 \tau^2)^{-(\nu + \frac{d+1}{2})} d\omega d\tau \\
&= O(1) \qquad \text{for } d > 2\,\nu > 0
\end{aligned}$$

■

**Lemma 2.11.7.** *$lim_{P,T \to \infty, \infty} \lambda_{max} \Sigma(\theta^*) \circ K_{Taper} = O(1)$ and $lim_{P,T \to \infty, \infty} \lambda_{min} \Sigma(\theta^*) \circ K_{Taper}$*

*Proof.* We use the result that

$$\min K_{Taper\,i,i} \lambda_{min} \Sigma(\theta^*) \leq \lambda_{max} \Sigma(\theta^*) \circ K_{Taper} \leq \max K_{Taper\,i,i} \lambda_{max} \Sigma(\theta^*)$$

■

*Proof.* By Taylor series expansion, $\Phi(X) = \Phi(y) + (X-y)\phi(y) + o_p(X-y)$ holds for almost surely a random variable $X$

Now taking $X = \frac{\hat{\Delta}^T\Sigma^{-1}\hat{\Delta}}{2\sqrt{\hat{\Delta}^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\Delta}}}$ and $y = \frac{\sqrt{\Delta^T\Sigma^{-1}\Delta}}{2}$.

Notice that $\hat{\Delta}^T\Sigma^{-1}\hat{\Delta} \sim F(\tilde{P}_T, n-2)$ with non central parameter since $\hat{\Delta}$ and $\hat{\Sigma}$ are independent

■

If we use weighted mixture of Matern Covariance function,we can show that all of the above conditions hold for covariance matrix $\Sigma_{PT\times PT}(\theta)$

**Lemma 2.11.8.** The expression of derivative of covariance function with respect to parameter $\theta$ is given below

*Proof.* By Fubini's theorem, we know that the differentiation and integral can be exchanged.

$$
\begin{aligned}
&\frac{\partial\gamma(s,t;\theta)}{\partial\theta} \\
&= K(\theta) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}} exp(\omega's+\tau't)\frac{\partial f(\omega,\tau;\theta)}{\partial\theta}d\omega d\tau \\
&= -(\nu+\frac{d+1}{2})\int_{\mathbb{R}^d}\int_{\mathbb{R}} exp(\omega's+\tau't)g(\omega,\tau)^{-(1+\nu+\frac{d+1}{2})} \\
&\begin{pmatrix}
2\alpha\eta(\beta^2+\tau^2) \\
2\beta\eta(\alpha^2+\tau^2) \\
\eta(\|\omega\|^2\tau^2) \\
(\alpha^2\beta^2+\beta^2\|\omega\|^2+\alpha^2\tau^2+\epsilon\|\omega\|^2\tau^2) \\
\frac{log(\alpha^2\beta^2+\beta^2\|\omega\|^2+\alpha^2\tau^2+\epsilon\|\omega\|^2\tau^2)}{\nu+\frac{d+1}{2}}
\end{pmatrix} d\omega d\tau
\end{aligned}
$$

We can verify that each of the element of $K(\theta)$ has finite integral multiplied with a continuous function of $\theta$. Since $\Theta$ is compact space in $\mathbb{R}^q$, $K(\theta)$ is bounded . ■

**Lemma 2.11.9.** The maximum eigenvalue of matrix $\frac{\partial\Sigma}{\partial\theta}$ is bounded from above

*Proof.* We will show that $\|\frac{\partial\Sigma}{\partial\theta}\|_2 \leq \|\frac{\partial\Sigma}{\partial\theta}\|_1 = \int_{\mathbb{R}^d}\int_{\mathbb{R}} \frac{\partial\gamma(s,t;\theta)}{\partial\theta} < \infty$ ■

**Lemma 2.11.10.** The maximum eigenvalue of matrix $\frac{\partial^2\Sigma}{\partial\theta^2}$ is bounded from above

*Proof.* Similar to proof above either we can differentiate again and apply Fubini Theorem ■

**Lemma 2.11.11.** $\|\Sigma\|_F^2 = O_pPT >> O_p(PT)^{\frac{1}{2}}$

*Proof.* Since covarince function is isotropic and stationary $\|\Sigma\|_F^2 = \sum_{i,j}\gamma^2(h_{ij};\theta)$

$PT\sum_i \min_j \gamma^2(h_{ij};\theta \leq \sum_{i,j}\gamma^2(h_{ij};\theta) \leq PT\sum_i \max_j \gamma^2(h_{ij};\theta)$

Using the fact that $\sum_i \max_j \gamma^2(h_{ij};\theta) \leq \int_0^\infty \gamma^2(x;\theta)dx < \infty$ ■

**Lemma 2.11.12.** If $\lambda_1 \geq \lambda_2 ... \geq \lambda_n$ be eigenvalues of positive semi definite matrix the condition number $k_n = \frac{\lambda_{max}}{\lambda_{min}}$can be have upper bound $1 + \frac{2v(A)\sqrt{(n-1)}}{m(A)-v(A)\sqrt{(n-1)}}$ and lower bound $1 + \frac{2v(A)/\sqrt{(n-1)}}{m(A)-v(A)/\sqrt{(n-1)}}$

where $m(A) = tr(A)/n$ and $v(A) = tr(A^2)/n - m^2$

*Proof.* From lemma in Wolkowicz & Styan (1980) ■

**Lemma 2.11.13.** Assuming general condition 13 holds, we have the results

- $\|\Sigma_t - \Sigma_{t,Taper}\|_1 = O_p(\frac{1}{\sqrt{P_t}})$
- $\|\Sigma_{k,t} - \Sigma_{k,t,Taper}\|_1 = O_p(\frac{1}{\sqrt{P_t}})$
- $\|\Sigma_{k,j,t} - \Sigma_{k,j,t,Taper}\|_1 = O_p(\frac{1}{\sqrt{P_t}})$

*Proof.* $\|\Sigma_t - \Sigma_{t,Taper}\|_1 \leq 2\frac{K_t\rho}{w_p}\int_{w_p}^\infty x^d|\gamma(x:\theta)|dx$ ■

Similar result holds for derivative and double derivative of $\Sigma(\theta)$

proof for theorem 7.2

*Proof.* By using Lemma 8.1, 8.3 ,7.1 we have established all regularity conditions for tapered matrix $\tilde{\Sigma}_{Tapered}$ $W(\hat{\beta}_{PMLE}) = 1 - \Phi(\frac{\Delta^T\Sigma^{-1}\Delta)(1+O_p(\frac{1}{\sqrt{n\tilde{P}_T}})+\frac{\tilde{P}_T}{n}(n_1-n_2)(1+O_p(\frac{1}{\sqrt{n\tilde{P}_T}}))}{2\sqrt{\Delta^T\Sigma^{-1}\Delta)(1+O_p(\frac{1}{\sqrt{n\tilde{P}_T}})+\frac{\tilde{P}_T}{n}(n_1+n_2)(1+O_p(\frac{1}{\sqrt{n\tilde{P}_T}}))}}$

■

# Chapter 3

# Random Projection for Tensor data

## 3.1 Introduction

Given a vector $x \in \mathbb{R}^{p\times 1}$, we can project it to lower dimensional space through linear map $f : \mathbb{R}^{p\times 1} \to \mathbb{R}^{d\times 1}$ defined by

$$f(x) = \frac{1}{\sqrt{d}^k}\tilde{A}^{(\mathcal{K})}x$$

where $\tilde{A}_{p\times d}$ is random matrix formed with element sampled independently from special classes of distributions.

$$\tilde{a}_{i,j} \overset{iid}{\sim}_{\mathcal{D}} a$$

This linear projection also preserves pairwise distances with high probability. This phenomenon is due to JL lemma for some $\epsilon > 0$) and constant $C$ independent of dimension.For $k = 1$

$$\mathbb{P}(1-\epsilon \leq \frac{\|f(x) - f(x')\|_2^2}{\|x - x'\|_2^2} \leq 1-\epsilon) \leq 2\exp(-\frac{d}{8}(\epsilon^2/2 - \epsilon^3/3))$$

**Lemma 3.1.1.**

$$\mathbf{P}(< f(x), f(y) > - < x, y > \geq \epsilon < x, y >) \leq 2 \sup_{x \in \mathbb{R}^p} \mathbf{P}(\|f(x)\|^2 - \|x\|^2 \geq \epsilon \|x\|^2)$$

*Proof.* Sun et al. (2018) lemma A.1 ■

Suppose there are n points in $\mathbb{R}^{p \times 1}$ corresponding to data matrix $X = [X_1..X_i..X_n]_{1 \leq i \leq n}$. We can chose $d \geq 2\log(n)\epsilon^{-2}$ such that all pairwise distances belonging to each of $\frac{n(n-1)}{2}$ pairs are preserved with high probability.

There an be various choices of distributions for random variable $a$. For example, $a \sim \mathbf{N}(0, 1)$ or in sparse RP $s = 3$ and very sparse RP $s = \sqrt{d}$ respectively

$$a_{i,j} = \begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

P. Li et al. (2006) $a_{i,j} = \begin{cases} +\sqrt{s} & \text{with probability } \frac{1}{2s} \\ 0 & \text{with probability } 1 - \frac{1}{s} \\ +\sqrt{s} & \text{with probability } \frac{1}{2s} \end{cases}$

## 3.2 Kronecker factors

Supposedly $\tilde{A}(\mathcal{K})_{p \times d}$ could be decomposed as Kronecker product of k matrices

$$\tilde{A}^{(\mathcal{K})} = A^{(1)} \otimes ..A^{(l)}.. \otimes A^{(k)}$$

where matrix $A^{(l)}$ is of order $d_l \times p_l$ for each $l \in \{1, 2, .., k\}$.Trivially $p = \prod_{l=1}^{k} p_l$ and $d^k$. The elements of product matrix are given by

$$\tilde{a}_{i,j} = a^{(1)}_{i_1,j_1} a^{(2)}_{i_2,j_2} .... a^{(l)}_{i_l,j_l} .. a^{(k)}_{i_k,j_k}$$

where $j = 1 + \sum_{l=1}^{k} (j_l - 1) \prod_{r=l+1}^{k} p_r$ and $i = 1 + \sum_{l=1}^{k} (i_l - 1) d^{k-l-1}$

For $l = k$ the expressions $\prod_{r=l+1}^{k} p_r$ and $\prod_{r=l+1}^{k} d_r$ assumed to be 1.

If we take each elements of the matrices $\{A^l\}_{1 \le l \le k}$ to be independently and identically distributed of each other, the resulting elements of product matrices $\tilde{A}_{i,j}$ are no longer mutually independent but product of independent random variables, thus weekly dependent in the sense it is $\rho$ mixing, and martingale with respect to filtration In this work, we show that even under this dependence structure we can achieve JL kind of for finite $p$ and $d$. We further assume that $\mathbb{E}(a^{(l)}_{i_l,j_l}) = 0$ and $\mathbb{E}[(a^{(l)}_{i_l,j_l})]^2 = 1$ for all $l \in [k], j \in [p_l], i \in [d_l]$.

### 3.2.1 Literature reviews

Concentration inequalities for quadratic forms like ours is known as Hansen Wright inequality.For dependent variables these type of inequality are discussed in Adamczak et al. (2015) which certain concentration property for 1 Lipschitz function which is hard to verify in our cases.Samson et al. (2000) discussed the concentration property for 1 Lipschitz function for strong mixing sequence and Markov chain.However strong mixing conditions like $\phi$ mixing is also difficult to verify Another approach would be use Herbst argument applicable to Log Sobolev measure as mentioned in Ledoux (1999).But proving such inequality is much more tedious in our case

Define a random variable $Y_i(x) =< \tilde{A}_i^{\mathcal{K}}, x >$ ,our reader may notice that for fixed $i$, $Y_i = \sum_{j_1=1,j_2=1,...j_K=1}^{(p_1,p_2,...p_K)} a_{i_1,j_1}^{(1)} a_{i_2,j_2}^{(2)} ....a_{i_l,j_l}^{(l)} ..a_{i_k,j_k}^{(k)} x_j$

given $j = 1 + \sum_{l=1}^{k} (j_l - 1) \prod_{r=l+1}^{k} p_r$ form a martingale w.r.t filtration $\mathcal{F}_l$ and $1 \leq l \leq K$. But concentration inequalities for martingale differences provides very poor bounds.

Readers may also notice stationarity a random variables $Y_i$ although actual distribution is intractable. $Y_i$ is popularly known as Polynomial Gaussian Chaos in probability literatures. Statisticians recognize this expression as general U statistics. Its moments and tail bounds are widely studied in Adamczak et al. (2015) and Latała et al. (2006). But in above literatures,moment bound involve supreme of empirical processes which are hard to estimate. Also, exponential bounds given in these literature is depended upon various semi norms of vector $x$ based partition of subset of $\{1, 2, ...K\}$. But it is desirable to derive inequalities in terms of $L_2$ norm of $x$ only.

Taking all the above bottleneck into consideration, we follow approach by Schudy & Sviridenko (2012) where moment bounds are calculated through brute force combinatorial argument. To our aid we have result known as hyper- contractivity which provides bound of $\|Y\|_r$ through $(\|Y\|_2)^r$ in Janson et al. (1997) theorem 5.10 and theorem 6.7. Below is the result stated

## 3.3 Weak dependence

By convention, for a random variable $X \in (\mathbb{R}, \mathcal{B}_R, \mathbf{P})$, $\|X\|_p = (\mathbb{E}(|X|))$. Define filtration sigma filed $\mathcal{F}_K = \sigma(A_{i_l,j_l}^{(l)} : d \leq i_l \leq 1, p_l \leq j_l \leq 1; K \leq l \leq 1)$

Suppose there is a sequence of random variables $(\tilde{a}_i)_{i=1}^n$ and $\tilde{a}_n \in (\mathbb{R}, \mathcal{B}_R, \mathbf{P})$,define sigma field $G_1^k = \sigma(\tilde{a}_i : 1 \leq i \leq h)$ rho $\rho(n) = \sup_{\substack{k \geq 1 \\ \tilde{a} \in G_1^h \\ \tilde{b} \in G_{h+n}^{\infty}}} |Cov(\tilde{a}, \tilde{b})| / \|\tilde{a}\|_2 \|\tilde{b}\|_2$ a sequence is $\rho$ mixing if

$\lim_{n \to \infty} \rho(n) \to 0$

In our case, $\rho(|i-i'|) = \begin{cases} 0 & \text{for } a \sim \mathbf{N}(0,1) \\ 0 & \text{for } a \sim \text{very sparse with } s = \sqrt{d} \end{cases}$

for $i \neq i'$

*Proof.* Define $Y_i(x) =< \tilde{A}_i^{\mathcal{K}}, x >$. for $i \neq i'$ but sometimes $j = j'$

$$\mathbb{E}(Y_i, Y_i') = \mathbb{E}(\sum_j \sum_{j'} \tilde{a}_{i,j'} \tilde{a}_{i',j} x_j x_{j'})$$

Now, for $i \neq i'$ with $i_l = i_l'$ but $i_k = i_k'$ without loss of generality and $j = j'$ $\mathbb{E}(\tilde{a}_{i,j}\tilde{a}_{i',j'})$

$= \mathbb{E}(a^{(1)}_{i_1,j_1} a^{(2)}_{i_2,j_2} ....a^{(l)}_{i_l,j_l} ..a^{(k)}_{i_k,j_k} a^{(1)}_{i_1',j_1} a^{(2)}_{i_2',j_2} ....a^{(l)}_{i_l',j_l} ..a^{(k)}_{i_k',j_k})$

$= \mathbb{E}(a^{2\,(1)}_{i_1,j_1} a^{2\,(2)}_{i_2,j_2} ....a^{2\,(l)}_{i_l,j_l} ..a^{(k)}_{i_k,j_k} a^{(k)}_{i_k',j_k}) = 1\,\mathbb{E}(a^{(k)}_{i_k,j_k})\mathbb{E}(a^{(k)}_{i_k',j_k}) = 0$

Due to independence. Hence whole expression has expectation 0. Therefore covariance is 0. ■

## 3.4 Concentration Inequality

For random variables following Gaussian distribution and very sparse distribution P. Li et al. (2006), we are able to prove this bound. We require some concentration bound for expression

$$\mathbf{P}(|\frac{1}{d^k} \frac{x^T \tilde{A}^{(\mathcal{K}),T} \tilde{A}^{(\mathcal{K})} x}{\|x\|_2^2} - 1| > t)$$

to decay at exponential else we would not attain the efficiency like JL lemma meaning the lower projected dimension will be much more than standard result for k=1,i.e we want $d = O(\log n)$.

**Theorem 3.4.1.** JL lemma for Gaussian or Raedmacher For some constant $m = m(K)$ depending on $K$

$$\mathbf{P}(|\frac{1}{d^k} \frac{S_{d^k}}{\|x\|_2^2} - 1| > t) < \frac{e^{-\frac{d}{m(K)} t^{\frac{2}{K}}}}{m(K)}$$

*Proof.* Since $\| \frac{S_{d^k}}{\|x\|_2 d^k} \|_{2r}^{2r} = O(\| \frac{N_1}{d} \frac{N_2}{d} .. \frac{N_k}{d} \|_{2r}^{2r})$ where $N_q \overset{iid}{\sim} \mathbb{N}(0,1)$ . Then it follows that $\mathbb{E}(e^{t \frac{S_{d^k}}{\|x\|_2 d^k}}) < C\mathbb{E}(e^{t \frac{N_1}{d} \frac{N_2}{d} .. \frac{N_k}{d}})$ And using the result stated in Achlioptas (2003), we complete the proof. Also note that this inequality is consistent with Latała et al. (2006). ■

### 3.4.1 Choice of d

We choose $d > \frac{m(K) log n}{t^{2/K}}$ using union bound so that distance between n points are preserved.

### 3.4.2 Memory efficiency

Under Kronecker decomposition, we need to store $\sum_{l=1}^{k} p_l d$ elements as compared to original matrix $\prod_{l=1}^{k} p_l d^k$, this is huge reduction even after considering trade-off in the probabilistic bound.

#### 3.4.2.1 Tensor type data

Another application can be for tensor data $x \in \mathbb{R}^{p_1 \times p_2 \times .. p_k}$ we can also obtain lower dimensional embedding which preserves distance with high probability.

$$y_{d_1 \times d_2 \times .. d_k} = \frac{1}{\sqrt{\prod_{l=1}^{k} d_l}} (A^{(1)} \otimes .. A^{(l)} .. \otimes A^{(k)})^T \, x_{p_1 \times p_2 \times .. p_k}$$

$$Vec(y)_{d \times 1} = \frac{1}{\sqrt{d}} (A^{(1)} \otimes .. A^{(l)} .. \otimes A^{(k)})^T \, Vec(x)_{p \times 1}$$

### 3.4.3 Variance reduction through averaging

Since all random projection are non adaptive methods. In most of the literature, iit is recommended that we generate several independent RPs and take their average. This ensemble technique also provide reduction in variance.

Let $\{\tilde{A}^h\}_{1 \leq h \leq H}$ be H independent copies of RP big Kronecker matrices. WE can define new

ensemble RP as

$$f_{*,H}(x) = \frac{1}{\sqrt{H}} \sum_{h=1}^{H} \tilde{A}^h x$$

## 3.5 Simulation result

Our first experiment evaluates the quality of the isometry for maps We generate n = 10 independent vectors $x1; ..; xn$ of sizes d = 2500; 10000; 40000 .We consider the following three RPs: 1. Gaussian RP; 2. Sparse RP ; 3. Very Sparse RP . For each, we compare the performance of RP, KRP with order 4 and d1 = d2=d3=d4. We evaluate the methods by repeatedly generating a RP and computing the reduced vector, and plot the ratio of the pairwise distance 1.

| Method | Gaussian | Sparse | Very Sparse |
|---|---|---|---|
| RP | 0.1409 (0.0015) | 0.1407 (0.0013) | 0.1412 (00.0014) |
| KRP(4) | 0.1431 (0.0016) | 0.1431 (0.0015) | 0.1520 (0.0033) |

Table 3.1: Average of total deviation of ratios of pairwise distance between projected and actual data from 1 . (Variability)

## 3.6 Future scope

We believe that such bounds can be achieved for wider class of random variables following identity $\|Y\|_r^r \leq (r)M\|Y\|_{r-1}^{r-1}$ for all $r$ as shown in Schudy & Sviridenko (2012).

## 3.7 Introduction to Tensors

With the advancement of information and engineering technology, modern days data science problems often come with gigantic size and increased complexity. On popular technique of storing

these data is use of multi-dimensional arrays, which preserves the data anatomy and contain multi-dimensional spatial and spatio-temporal correlations.

A tensor is a multi-dimensional or d-way array, which is a generalization of data matrix in a higher dimensional situation. More formally, according to Kolda & Bader (2009) and Hackbusch (2012), a d-way tensor is an element of the tensor space generated by the tensor products of d vector spaces. Similar to the traditional vector based machine learning literature, the learning of tensor can also be divided into supervised learning and unsupervised learning. Unsupervised tensor learning generally involves the tensor decomposition and feature selection. Some theoretical results and applications about unsupervised tensor learning can be found in Kolda & Bader (2009), X. He et al. (2006), Chi & Kolda (2012), De Lathauwer et al. (2000), and Lu et al. (2008). The framework of supervised tensor learning has been proposed by Tao et al. (2005), in which one can learn a tensor based rule from training data for classification and regression. The tensor regression problem has been widely studied. Such examples include Zhou et al. (2013), Wimalawarne et al. (2016), L. Li & Zhang (2017), Lock (2018), and Raskutti et al. (2019). As an indispensable part of the supervised tensor learning problem, however, the tensor type data classification is under developed. Research on tensor classification includes Zhou et al. (2013), Pan et al. (2018), Signoretto et al. (2011), L. He et al. (2014), Q. Li & Schonfeld (2014), and Tan et al. (2012).

### 3.7.1 Limitation for Gaussian assumptions

There are several major deficiencies under the current development. Firstly, the assumption about data distribution. The Gaussian assumption for tensor type of data, e.g., Pan et al. (2018) may not be adequate in many applications. Since the probability theory for tensor data has not been well established, probabilistic discriminant such as LDA and Bayes classifier may have limitations from theoretical foundation perspective. Secondly, the distance based methods, e.g., Lu et al. (2008) and Q. Li & Schonfeld (2014). These methods train classifier based on tensor Frobenius norm, which is

not suitable for high dimensional tensor data. It is known that the high dimensional data has issues with L2 norm and Frobenius norm. For example Domingos (2012) wrote " If we approximate a hypersphere by inscribing it in a hypercube, in high dimensions almost all the volume of the hypercube is outside the hypersphere". Beyer et al. (1999) showed that the difference between the maximum and minimum distances to a given query point does not increase as fast as the nearest distance to any point in high dimensional space. As data dimension increases, these distance based method may fail. Signoretto et al. (2011) and L. He et al. (2014) performed classification with kernels however, there is no theoretical results about classification error established.

## 3.8 Preliminaries

### 3.8.1 Mathematical Background for Tensor

We first introduce standard tensor notation and operations (e.g. Kolda & Bader (2009)) that are used in this paper. Numbers and scalars are denoted by lowercase letters such as $x, y$. Vectors are denoted by boldface lowercase letters, e.g. $\mathbf{a}$. Matrices are denoted by boldface capital letters, e.g. $\mathbf{A}, \mathbf{B}$. A higher-dimensional tensor is a generalization of vector and matrix representation for higher order data, which are denoted by boldface Euler script letters such as $\mathcal{X}, \mathcal{Y}$. Notations for vector and tensor spaces will be sepcified when necessary.

The order of a tensor is the number of dimensions, also known as ways or modes. For example, a scalar can be regarded as a order zero tensor, a vector can be a order one tensor, and a matrix can be a order two tensor. In general, a tensor can have $d$ modes as long as $d$ is an integer.

The way of indicating entries of tensors is same as we do for vectors and matrices. The $i$-th entry of a vector $\mathbf{x}$ is $x_i$, the $(i, j)$-th element of a matrix $\mathbf{X}$ is $x_{i,j}$, and the $(i_1, ..., i_d)$-th element of a d-way tensor $\mathcal{X}$ is $x_{i_1,...,i_d}$. The indices of a tensor $i_1, ..., i_d$ range from 1 to their captial version, e.g. $i_k = 1, ...., I_k$ for every mode $k = 1, ...d$.

Sub-arrays of a tensor are formed when a subset of the indices are fixed. Similar to matrices that have rows and columns, high-dimensional tensors have various types of sub-arrays. For example, by fixing every index but one in a d-way tensor, we can get one of its fibers, which are analogue of matrix rows and columns. Another type of frequently used tensor sub-arrays is slice, which is a two dimensional section of a tensor. A slice of a tensor can be defined by fixing all but two indices. We will use $\mathbf{x}_{:i_2\ldots i_d}$ to denote one fiber of a d-way tensor, and use $\mathbf{X}_{::i_3\ldots i_d}$ to denote one of its slices.

## 3.9 Kernelized Support Tensor Machine

### 3.9.1 Framework of the Classification Problem

The classification problem for tensor data is a problem of learning a tensor from the training data. Let $T = \{(\mathcal{X}_1, y_1), ...., (\mathcal{X}_n, y_n)\}$ be the training set, where $\mathcal{X}_i \in \mathbb{R}^{I_1 \times I_2 \ldots \times I_d}$ are d-mode tensors, $y_i$ are labels. If we assume the training risk of a classifier $f \in \mathcal{X}^*$ is $\hat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} I(f(\mathcal{X}_i) \neq y_i)$, the problem will be looking for a $\hat{f}$

$$\hat{f} = \{f : \hat{\mathcal{R}}_n(f) = \min \hat{\mathcal{R}}_n(f), f \in \mathcal{H}^*\} \tag{3.1}$$

To solve this problem, we need to search all functions in the functional tensor space, which is challenging. However, the problem will be simplified if the function is an universal reproducing kernel. This will let us find the solution only on the Hilbert space embedded by continuous functions. We shall prove this claim by presenting a represent theorem for kernelized support tensor machine. Further, we would be able to prove the consistency of our classifier.

### 3.9.2 Support Tensor Machine

Generalizing the support vector machine to tensor version is quite straightforward. The kernelized support tensor machine classifier is

$$sign(f(\mathcal{X})) \tag{3.2}$$

The function $f$ is the optimal of the following objective function

$$\min_{f} \quad \lambda||f||_* + \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathcal{X}_i)) \tag{3.3}$$

where $||f||_*$ is the norm of the functional tensor space $\mathcal{H}^*$, and $L$ is a measurable loss function defined on $(\mathcal{X}^* \times Y)$. $\mathcal{X}^*$ is an algebraic tensor space. The optimal function is a measureable function $f : \mathcal{X}^* \to \mathbb{R}$, where $\mathcal{X}^* = \mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2} ... \otimes \mathbb{R}^{I_d}$.

The representer theorem in support vector machine says the solution of support vector machine can be written as a linear combination of kernel functions. As a result, one only needs to learn the coefficients in the linear combination instead of learning the whole function. Similarly, we can also propose a representer theorem for this tensor learning problem.

**Theorem 3.9.1. (Tensor Representer Theorem)**

Let $K(\cdot, \cdot)$ be a fixed kernel coming from the $\mathcal{K}(\mathcal{X}^* \times \mathcal{X}^*, \mathbb{R})$, $\mathcal{H}^*$ be the corresponding Reproducing Kernel Hilbert Space. Let $L$ be an arbitrary loss function. If the optimization function (3.3) has optimal solutions, then all the solutions can be written in the following way:

$$\hat{f}(\mathcal{X}) = \sum_{i=1}^{n} \hat{\beta}_i K(\mathcal{X}_i, \mathcal{X})$$

The proof of this theorem is attached in our appendix. As a direct benefit of this result, one just needs to estimate those parameters $\hat{\beta}_i$ in order to get a optimal solution.

### 3.9.3 STM with random projection

We can choose $K^{(j)} = A^{(j)}$ where $A^{(j)}$ is random projection matrix thus

$$\mathbf{K} = (\sum_{k,l=1}^{r}\prod_{j=1}^{d} A^{(j)}(\mathbf{x}_{k,1}^{(j)}, \mathbf{x}_{l}^{(j)}), ..., \sum_{k,l=1}^{r}\prod_{j=1}^{d} A^{(j)}(\mathbf{x}_{k,n}^{(j)}, \mathbf{x}_{l}^{(j)}))^T$$

. $r$ is the CP rank of the algebraic tensor space, $\mathbf{x}_{ik}^{(j)}$

### 3.9.4 Solving the STM

Now we start discussing the way of estimating our Support tensor machine from a group of training tensor and their corresponding labels. We are going to consider only the cumulant-based tensor kernel functions introduced in the previous section in this part. The situation of naive tensor kernels are straightforward, and one can use traditional SVM method to estimate with only a slight modification. According to the representer theorem and the definition of cumulant-based kernel function, we assume that the solution of support tensor machine has the following explicit form.

$$\begin{aligned}\hat{f}(\mathcal{X}) &= \sum_{i=1}^{n}\beta_i \sum_{k,l=1}^{r}\prod_{j=1}^{d} K^{(j)}(\mathbf{x}_{k,i}^{(j)}, \mathbf{x}_{l}^{(j)}) \\ &= \mathbf{K}^T\beta\end{aligned} \tag{3.4}$$

where $\beta = (\beta_1, ..., \beta_n)^T$ and $\mathbf{K} = (\sum_{k,l=1}^{r}\prod_{j=1}^{d} K^{(j)}(\mathbf{x}_{k,1}^{(j)}, \mathbf{x}_{l}^{(j)}), ..., \sum_{k,l=1}^{r}\prod_{j=1}^{d} K^{(j)}(\mathbf{x}_{k,n}^{(j)}, \mathbf{x}_{l}^{(j)}))^T$. $r$ is the CP rank of the algebraic tensor space, $\mathbf{x}_{ik}^{(j)}$ and $\mathbf{x}_{k}^{(j)}$ are components of tensor CP decomposition of the corresponding training and testing tensors. If the data do not come in CP form, we may need to perform a CP decomposition at first. Plugging the soluction (3.4) into the objection function

(3.3), we can get

$$\min_{\beta} \quad \lambda\beta^T \mathbf{K}\beta + \frac{1}{n}\sum_{i=1}^{n} L(y_i, \mathbf{K}^T(:,i)\beta) \tag{3.5}$$

where $\mathbf{K}(:,i)$ is the ith column of matrix $\mathbf{K} = [K_1, ..., K_n]$.

This problem can be solved directly with gradient descent. All values except the primal vector can be easily evaluated from the training set. The derivative of $\beta$ is

$$2\lambda\mathbf{K}\beta + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial L}{\partial \beta} \tag{3.6}$$

Let equation (3.6) equals to zero and solve for our problem. In our application, we took squared hing loss which $L(y, \mathcal{X}) = [\max(0, 1 - yf(\mathcal{X}))]^2$. The algorithm of training and prediction are described below denoted as algorithm **??** and algorithm **??** respectively: In algorithm **??**, the complexity of the training process is $O(n^2r^2\sum_{j=1}^{d} I_j)$, which is will be much smaller than the complexity of any vectorized method, $O(n^2\prod_{j=1}^{d} I_j)$, with low rank assumption. In addition, we suggest to save the decomposed list of the training data when handling high dimensional data in algorithm **??**. This is because the decomposed list is much smaller than the original data, which saves memory for computers. Having decomposed list instead of the original data can also make the prediction faster, since one does not need to repeat the decomposition again.

### 3.9.5 Estimation with Complete Tensor Data

As we mentioned above, a CP decomposition is required when getting data ready for our algorithm. We can estimate components of a rank r tensor from this procedure, and can feed this estimation to our model. The estimated tensor decomposition will be an unique approximation for the original data under most situation. Due to this uniqueness, we take the proposition from Kolda & Bader (2009):

**Proposition 1.** *Let $\mathcal{X}$ be a d-way tensor with rank r. If it can be expressed as $\mathcal{X} = \sum_{k=1}^{r} \mathbf{x}_k^{(1)} \otimes \mathbf{x}_k^{(2)} ... \otimes \mathbf{x}_k^{(d)} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(d)}]$, where the columns of each $\mathbf{X}^j \in \mathbb{R}^{I_j \times r}, j = 1, ..., d$ are $X_k^{(j)} k = 1, ..., r; j = 1, .., d$ from the expression. The CP decomposition of this tensor is unique if*

$$\sum_{k=1}^{d} R(\mathbf{X}^{(j)}) \geqslant 2r + d - 1 \tag{3.7}$$

*where $R(\mathbf{X}^{(j)})$ are the corresponding column ranks of matrices $\mathbf{X}^{(j)}$. As a result, the probability of miss-classification is identical, i.e.*

$$\mathbb{P}(\hat{y} \neq y|\mathcal{X}) = \mathbb{P}(\hat{y} \neq y|\hat{\mathcal{X}}) \tag{3.8}$$

*where $\hat{\mathcal{X}}$ is the tensor with estimated CP decomposition*

For more details of the uniqueness of the decomposition, we refer Kolda & Bader (2009) and Sidiropoulos & Bro (2000). During our application, we use the Alternating Least Square (ALS) to estimate the decomposition.

The second issue about our model is the rank assumption. We assume all tensor data are ideally from the same tensor space with rank r. Under most situations, we do not know the rank $r$ apriori. We prescribe the ranke r when we pre-process the data for training by finding one which can provide the best approximation for tensor decomposition.

## 3.10 Statistical Property of STM

In the last part of our theoretical results, we want to highlight the performance and the generalization ability of our classifier for tensors. In the general evaluation of a decision rule, one will be interested in exploring the bound for its classification risk. For example, assume our data for classification are $\{(x, y) \in \mathcal{X}^* \times \mathcal{Y}\}$. Let $f$ be a decision rule for data generated from $\mathcal{X}^* \times \mathcal{Y}$, the

classification risk of this rule is defined as:

$$\mathcal{R}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \mathbf{Pr}(f(x) \neq y|x)\mu(dx)\mu(dy) \tag{3.9}$$

where $\mu$ are measures defined on $\mathcal{X}$ and $\mathcal{Y}$. The lower bound of this risk is the Bayes risk, which we denote with $\mathcal{R}^*$. For simplicity, we consider the binary classification case where $\mathcal{Y} = \{-1, 1\}$. In addition, we assume there is no noise in the problem such that $\forall \mathcal{X} \in \mathcal{X}^*$,

$$\mathbf{Pr}(f(\mathcal{X} = 1)|\mathcal{X}) \neq \mathbf{Pr}(f(\mathcal{X} = -1)|\mathcal{X}) \tag{3.10}$$

In other words, we will not consider the situation where posterior distribution does not provide a decision and one can only guess randomly.

However, a classification rule learned directly from a given training set usually will not be able to reach the Bayes risk. In fact, we can only estimate the empirical risk for a rule when an observation with length $n$ is given, which is

$$\hat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{f(x_i) \neq y_i\} \tag{3.11}$$

$\{(x_i, y_i), i = 1, ..., n\}$ is our training set. We always try to minimize this empirical risk with some methods, and find out the optimal classifier $\hat{f}$ for this risk. This is what we called Empirical Risk Minimization(ERM) in general statistical learning problems. The evaluation of the method we followed to find $\hat{f}$ depends on the bound

$$\delta_n = |\mathcal{R}^* - \mathcal{R}(\hat{f})| \tag{3.12}$$

If this quantity $\delta$ converges to zero as the sample size increase, then the classifier is consistent and

the method we follow is statistically sound. This quantity can be bounded by

$$\delta_n \leqslant |\mathcal{R}^* - \mathcal{R}^*(f)| + |\mathcal{R}^*(f) - \mathcal{R}(\hat{f})| \tag{3.13}$$

where $\mathcal{R}^*(f) = \inf_{f \in \mathcal{H}} \mathcal{R}(f)$ is the minimal risk of a collection of classifiers $\mathcal{H}$. According to the result from Steinwart & Christmann (2008), we have the following theorem:

**Theorem 3.10.1.** If $\mathbf{K}$ is a universal kernel on a compact subset of tensor space $\mathcal{X}^*$. The loss function $L$ is Lipschitz continuous. Then $\mathcal{R}^* = \mathcal{R}^*(f)$.

One can see Steinwart & Christmann (2008) for the proof of this theorem. The intuition of this theorem is pretty simple since we have shown the universal approximation property of tensor based kernel functions. This result will always hold in the classification problems since the loss functions such as hinge loss are always Lipschitz continuous. The consistency problem turns out to be finding bounds for $|\mathcal{R}^*(f) - \mathcal{R}(\hat{f})|$. Our next result shows the convergence of this part.

**Theorem 3.10.2.** Let $\mathbf{V}$ be a compact subspace of the algebraic tensor space $\mathcal{X}^*$. Let $\mathbf{K}$ be a cumulant-based kernel function for tensor that is universal on the $\mathbf{V}$, and $|\mathbf{K}|_\infty \leqslant 1$. If we assume the tensor space has dimension $p = I_1 \times I_2 \times ... \times I_d < \infty$. Then for all Borel probaility measure $\mathbf{Pr}$ on $(\mathcal{X}^* \times \mathcal{Y})$ satisfying $\mathbf{Pr}(f(\mathcal{X} = 1)|\mathcal{X}) \neq \mathbf{Pr}(f(\mathcal{X} = -1)|\mathcal{X})$, we have

$$\mathcal{R}(\hat{f}_n) \to \mathcal{R}^*(f) \qquad n \to \infty \tag{3.14}$$

in probability.

# APPENDIX

# APPENDIX

**Main lemmas**

**Lemma 3.10.3.** $\mathbb{E}(< \tilde{A}_i^{\mathcal{K}}, x >) = 0$ for any row $i$

**Lemma 3.10.4.** $\|(< \tilde{A}_i^{\mathcal{K}}, x >)\|_2 = \|x\|_2$ for any row $i$

*Proof.* Define $a^{(1)}_{i_1,j_1} a^{(2)}_{i_2,j_2} .... a^{(l)}_{i_l,j_l} .. a^{(k)}_{i_k,j_k} = \tilde{a}_{i,j}$ $\|(< \tilde{A}_i^{\mathcal{K}}, x >)\|_2^2 = \mathbb{E} \sum_{j=1}^{P} x_j^2 \, (\tilde{a}_{i,j})^2$

$+ \sum_{\substack{j'=1 \\ j \neq j'}}^{P} \sum_{j=1}^{P} x_j x'_j \, \tilde{a}_{i,j} \tilde{a}_{i,j'}$

Second term is zero due to conditional expectation property. ■

First we try to estimate the $r$ th moment of $S_d^2(x) = \frac{x^T \tilde{A}^{(\mathcal{K}),T} \tilde{A}^{(\mathcal{K})} x}{d^K} = \frac{\sum_{i=1}^{d} \|\tilde{A}_i^{(\mathcal{K})} x\|_2^2}{d^K}$ in terms $Y_1^2 = (\tilde{A}_1^{(\mathcal{K})} x)^2$, since $\{Y_i\}_{i=1}^{d^k}$ are exchangeable, stationary, weak $\rho$ mixing sequence. Unfortunately,

**Lemma 3.10.5.** Rosenthal Inequality for *rho* mixing sequences

For distribution following sparse distribution P. Li et al. (2006) Assume $\mathbb{E}(Y) = 0$ and $\|Y\|_{2r}^{2r} < \infty$ then there exists a positive constant $C = C_{r,\rho}$ such that

$$\mathbb{E} \max_{1 \leq i \leq d^k} (|S_{d^k}|)^{2r} \leq C(d^r \|x\|_2^{2r} + d(2r-1)^{rK} \|x\|_2^{2r})$$

*Proof.* Q.-M. Shao (1995) theorem 1, using $\rho = 0$ and $\|Y_1\|_2^2 = \|x\|_2^2$ and exchangeability of $Y_i$ now using result that $\|Y_1\|_{2r}^{2r} \leq (2r-1)^{rK} \|Y_1\|_2^{2r}\|$ ■

**Lemma 3.10.6.** Hypercontractivity for Gaussian distribution

Suppose $Y$ is $k$ degree Gaussian polynomial chaos in some Gaussian Hilbert space, then $\|S_d\|_{2r}^{2r} \leq (2r-1)^{Kr} \|S_d\|_2^{2r} = (r-1)^{K/2} d \|Y\|_2^2 = \|x\|_2^2$

*Proof.* Janson et al. (1997) theorem 5.10 and theorem 6.7 ■

**Lemma 3.10.7.** Hypercontractivity for Gaussian distribution

Suppose $Y$ is $k$ degree Gaussian polynomial chaos in some Gaussian Hilbert space, then $\|S_{d^k}\|_{2r}^{2r} \leq (2r-1)^{Kr}\|S_{d^k}\|_2^{2r} = (r-1)^{K/2}d\|Y\|_2^2 = \|x\|_2^2$

*Proof.* Janson et al. (1997) theorem 5.10 and theorem 6.7 ■

**Lemma 3.10.8.** Hypercontractivity for very sparse distribution $\|S_{d^k}\|_r \leq C_{r,K}\|S_{d^k}\|_2$

*Proof.* Janson et al. (1997) lemma 5.2 proves hypercontractivity for Raedmacher variable $Z$. Observe that for any constant $H$ , same will hold for $Z' = HZ$. Define a new random variable V independent of $Z'$such that $Pr(V = d^{1/4}) = \frac{1}{\sqrt{d}}$ and $Pr(V = 0) = 1 - \frac{1}{\sqrt{d}}$ , $a \sim VZ'$ any two point distribution is hyper contractive as well Janson et al. (1997) lemma 5.2. So, due to independence their product $VZ'$ is hyper contractive ■

# BIBLIOGRAPHY

# BIBLIOGRAPHY


Achlioptas, D. (2003). Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences*, *66*(4), 671–687.

Adamczak, R., et al. (2015). A note on the hanson-wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, *20*.

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., & Gee, J. C. (2011). A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, *54*(3), 2033–2044.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful? In *International conference on database theory* (pp. 217–235).

Bickel, P. J., Levina, E., et al. (2004). Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, *10*(6), 989–1010.

Cai, T., & Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, *106*(496), 1566–1577.

Chi, E. C., & Kolda, T. G. (2012). On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, *33*(4), 1272–1299.

Chu, T., Zhu, J., Wang, H., et al. (2011). Penalized maximum likelihood estimation and variable selection in geostatistics. *The Annals of Statistics*, *39*(5), 2607–2625.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fmri)"brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex. *Neuroimage*, *19*(2), 261–270.

Cressie, N., & Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, *94*(448), 1330–1339.

Cressie, N., & Lahiri, S. N. (1996). Asymptotics for reml estimation of spatial covariance parameters. *Journal of Statistical Planning and Inference*, *50*(3), 327–341.

De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, *21*(4), 1253–1278.

Domingos, P. M. (2012). A few useful things to know about machine learning. *Commun. acm*, *55*(10), 78–87.

Doshi, J., Erus, G., Ou, Y., Gaonkar, B., & Davatzikos, C. (2013). Multi-atlas skull-stripping. *Academic radiology*, *20*(12), 1566–1576.

Fan, J., & Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Annals of statistics*, *36*(6), 2605.

Fan, J., Feng, Y., & Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(4), 745–771.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348–1360.

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*(1), 101.

Furrer, R., Bachoc, F., & Du, J. (2016). Asymptotic properties of multivariate tapering for estimation and prediction. *Journal of Multivariate Analysis*, *149*, 177–191.

Furrer, R., Genton, M. G., & Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, *15*(3), 502–523.

Ghosh, D. (2001). Singular value decomposition regression models for classification of tumors from microarray experiments. In *Biocomputing 2002* (pp. 18–29). World Scientific.

Gutman, B. A., Hua, X., Rajagopalan, P., Chou, Y.-Y., Wang, Y., Yanovsky, I., ... others (2013). Maximizing power to track alzheimer's disease and mci progression by lda-based weighting of longitudinal ventricular surface features. *Neuroimage*, *70*, 386–401.

Hackbusch, W. (2012). *Tensor spaces and numerical tensor calculus* (Vol. 42). Springer Science & Business Media.

He, L., Kong, X., Yu, P. S., Yang, X., Ragin, A. B., & Hao, Z. (2014). Dusk: A dual structure-preserving kernel for supervised tensor learning with applications to neuroimages. In *Proceedings of the 2014 siam international conference on data mining* (pp. 127–135).

He, X., Cai, D., & Niyogi, P. (2006). Tensor subspace analysis. In *Advances in neural information processing systems* (pp. 499–506).

Hiai, F., & Lin, M. (2017). On an eigenvalue inequality involving the hadamard product. *Linear Algebra and its Applications*, *515*, 313–320.

Janson, S., et al. (1997). *Gaussian hilbert spaces* (Vol. 129). Cambridge university press.

Kaufman, C. G., Schervish, M. J., & Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, *103*(484), 1545–1555.

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, *51*(3), 455–500.

Kwon, S., & Kim, Y. (2012). Large sample properties of the scad-penalized maximum likelihood estimation on high dimensions. *Statistica Sinica*, 629–653.

Latała, R., et al. (2006). Estimates of moments and tails of gaussian chaoses. *The Annals of Probability*, *34*(6), 2315–2331.

Ledoux, M. (1999). Concentration of measure and logarithmic sobolev inequalities. In *Seminaire de probabilites xxxiii* (pp. 120–216). Springer.

Li, L., & Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association*, *112*(519), 1131–1146.

Li, P., Hastie, T. J., & Church, K. W. (2006). Very sparse random projections. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (pp. 287–296).

Li, Q., & Schonfeld, D. (2014). Multilinear discriminant analysis for higher-order tensor data classification. *IEEE transactions on pattern analysis and machine intelligence*, *36*(12), 2524–2537.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 13–22.

Lock, E. F. (2018). Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics*, *27*(3), 638–647.

Lu, H., Plataniotis, K. N., & Venetsanopoulos, A. N. (2008). Mpca: Multilinear principal component analysis of tensor objects. *IEEE transactions on Neural Networks*, *19*(1), 18–39.

Mai, Q., Zou, H., & Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, *99*(1), 29–42.

Matérn, B. (1960). Spatial variation, volume 36 of. *Lecture Notes in Statistics*.

Mourao-Miranda, J., Bokde, A. L., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: support vector machine on functional mri data. *NeuroImage*, *28*(4), 980–995.

Muschelli, J., Gherman, A., Fortin, J.-P., Avants, B., Whitcher, B., Clayden, J. D., ... Crainiceanu, C. M. (2018). Neuroconductor: an r platform for medical imaging analysis. *Biostatistics*.

Pan, Y., Mai, Q., & Zhang, X. (2018). Covariate-adjusted tensor classification in high dimensions. *Journal of the American Statistical Association*, 1–15.

Petersen, R. C., Aisen, P., Beckett, L. A., Donohue, M., Gamst, A., Harvey, D. J., ... others (2010). Alzheimer's disease neuroimaging initiative (adni): clinical characterization. *Neurology*, *74*(3), 201–209.

Popescu, V., Battaglini, M., Hoogstrate, W., Verfaillie, S. C., Sluimer, I., van Schijndel, R. A., ... others (2012). Optimizing parameter choice for fsl-brain extraction tool (bet) on 3d t1 images in multiple sclerosis. *Neuroimage*, *61*(4), 1484–1494.

Raskutti, G., Yuan, M., Chen, H., et al. (2019). Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, *47*(3), 1554–1584.

Rathi, V., & Palani, S. (2012). Brain tumor mri image classification with feature selection and extraction using linear discriminant analysis. *arXiv preprint arXiv:1208.2128*.

Rohlfing, T., Zahr, N. M., Sullivan, E. V., & Pfefferbaum, A. (2008). The sri24 multichannel brain atlas: construction and applications. In *Medical imaging 2008: Image processing* (Vol. 6914, p. 691409).

Samson, P.-M., et al. (2000). Concentration of measure inequalities for markov chains and $\phi$-mixing processes. *The Annals of Probability*, *28*(1), 416–461.

Schudy, W., & Sviridenko, M. (2012). Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the twenty-third annual acm-siam symposium on discrete algorithms* (pp. 437–446).

Shao, J., Wang, Y., Deng, X., Wang, S., et al. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of statistics*, *39*(2), 1241–1265.

Shao, Q.-M. (1995). Maximal inequalities for partial sums of $\rho$-mixing sequences. *The Annals of Probability*, 948–965.

Sidiropoulos, N. D., & Bro, R. (2000). On the uniqueness of multilinear decomposition of n-way arrays. *Journal of chemometrics*, *14*(3), 229–239.

Signoretto, M., De Lathauwer, L., & Suykens, J. A. (2011). A kernel-based framework to tensorial data analysis. *Neural networks*, *24*(8), 861–874.

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Sun, Y., Guo, Y., Tropp, J. A., & Udell, M. (2018). Tensor random projection for low memory dimension reduction. In *Neurips workshop on relational representation learning.*

Tan, X., Zhang, Y., Tang, S., Shao, J., Wu, F., & Zhuang, Y. (2012). Logistic tensor regression for classification. In *International conference on intelligent science and intelligent data engineering* (pp. 573–581).

Tao, D., Li, X., Hu, W., Maybank, S., & Wu, X. (2005). Supervised tensor learning. In *Data mining, fifth ieee international conference on* (pp. 8–pp).

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, *99*(10), 6567–6572.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, 104–117.

Varah, J. M. (1975). A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications*, *11*(1), 3–5.

Wang, Y., Fan, Y., Bhatt, P., & Davatzikos, C. (2010). High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables. *Neuroimage*, *50*(4), 1519–1535.

Wimalawarne, K., Tomioka, R., & Sugiyama, M. (2016). Theoretical and experimental analyses of tensor-based regression and classification. *Neural computation*, *28*(4), 686–715.

Witten, D. M., & Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(5), 753–772.

Wolkowicz, H., & Styan, G. P. (1980). Bounds for eigenvalues using traces. *Linear algebra and its applications*, *29*, 471–506.

Yasui, Y., & Lele, S. (1997). A regression method for spatial disease rates: an estimating function approach. *Journal of the American Statistical Association*, *92*(437), 21–32.

Yingjie, L., & Maiti, T. (2019). High dimensional discriminant analysis for spatially dependent data. *In Progress*.

Yu, H., & Yang, J. (2001). A direct lda algorithm for high-dimensional data—with application to face recognition. *Pattern recognition*, *34*(10), 2067–2070.

Yushkevich, P. A., Gao, Y., & Gerig, G. (2016). Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In *Engineering in medicine and biology society (embc), 2016 ieee 38th annual international conference of the* (pp. 3342–3345).

Zhou, H., Li, L., & Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, *108*(502), 540–552.

Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, *36*(4), 1509.