# Exploiting the Limits of Structure Learning via Inherent Symmetry

**Peng He**
Department of Automation
Tsinghua University

**Changshui Zhang**
Department of Automation
Tsinghua University

## Abstract

This theoretical paper is concerned with the structure learning limit for Gaussian Markov random fields from i.i.d. samples. The common strategy is applying the Fano method to a family of restricted ensembles. The efficiency of this method, however, depends crucially on selected restricted ensembles. To break through this limitation, we analyze the whole graph ensemble from a group theoretical viewpoint. The key ingredient of our approach is the invariance of orthogonal group actions on the symmetric Kullback-Leibler divergence. We then establish the connection of the learning limit and eigenvalues of concentration matrices, which further leads to a sharper structure learning limit. To our best knowledge, this is the first paper to consider the structure learning problem via inherent symmetries of the whole ensemble. Finally, our approach can be applicable to other graphical structure learning problems.

## 1 Introduction

Markov random fields (MRFs) offer a powerful tool for representing high-dimensional distributions and have found widespread applications in a variety of areas including computer vision (Geman and Geman, 1984), bio-informatics (Ahmedy et al., 2008)(Durbin et al., 1998) and social networks analysis (Fernando, 2007)(Wasserman and Faust, 1994). The problem of graphical model selection or structure learning refers to recover the unknown graph using observations from the underlying distribution. An active line of work imposes various restrictions on the underlying graph to obtain consistent procedures in the high-dimensional regime. A variety of methods have been proposed to estimate graph structures for general Markov random fields (Bresler et al., 2008), Ising models (Anadkumar et al., 2010) (Jalali et al., 2011)(Ravikumar et al., 2010) and Erdős-Rényi random graphs (Anadkumar et al., 2011).

For sparse Gaussian Markov random fields, a recent class of popular approaches have studied estimators based on the $\ell_1$-regularized Gaussian MLE (Graphical Lasso) methods (d'Aspremont et al., 2008)(Friedman et al., 2008)(Hsieh et al., 2011)(Ravikumar et al., 2008)(Yuan and Lin, 2007), which statistical guarantees are analyzed in (Ravikumar et al., 2008)(Rothman et al., 2008) under certain conditions on the covariance matrix. Another approach performs the linear neighborhood selection with $\ell_1$-regularization(Neighborhood Lasso) (Meinshausen and B$\ddot{u}$hlmann, 2006) using some regularity assumptions for the covariance matrices and neighborhood stability assumption, which sparsistency guarantees are investigated in (Meinshausen and B$\ddot{u}$hlmann, 2006)(Wainwright, 2009). More recently, greedy methods (Jalali et al., 2011)(Johnson et al., 2012) are applied to learn the graph structure the restricted eigenvalue and smoothness conditions. Furthermore, recent work has addressed information-theoretic limits of model selection inspired by techniques from information theory (Cover and Thomas, 1991) and nonparametric estimation (Yang and Barron, 1999)(Yu et al., 1996), establishing some necessary and sufficient conditions for model selection of Markov random fields (Anadkumar et al., 2011)(Bresler et al., 2008)(Wang et al., 2010) and Ising models (Anadkumar et al., 2010)(Santhanam and Wainwright, 2008).

In this paper, we consider the problem of the information-theoretic limit of model selection for Gaussian Markov random fields from i.i.d. samples, regardless of any algorithm and computational complexity. This problem is also equivalent to the estimation of the concentration matrix (inverse covariance matrix) of a zero-mean Gaussian random vector. We

Appearing in Proceedings of the $17^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

analyze this problem in the high-dimensional setting, where the graph size $p$ and the number of edges $k$ are allowed to scale with the sample size $n$. The common strategy resorts to Fano's inequality, which relies on estimating the mutual information between the underlying graphical models in restricted ensembles. And the mutual information is usually evaluated by upper bounds on it such as symmetrized Kullback-Leibler divergences (Wang et al., 2010) between graphs in the restricted ensemble. The main limitation of this method is that it is hard to evaluate symmetric (symmetrized) Kullback-Leibler divergences of generic graphs. Therefore, only some restricted ensembles such as $d$-clique graphs are analyzed in previous literature, limiting the strength of results obtained.

To overcome the above difficulty due to the selection of restricted ensembles, we directly handle the whole graph ensemble from a group theoretical perspective. More explicitly, we first investigate the geometry of concentration matrices. We find that concentration matrices of any two graphical models are connected by a space reflection or rotation in Euclidean space. Then we demonstrate the invariance of orthogonal group actions on concentration matrices and the symmetric Kullback-Leibler divergence. The symmetric Kullback-Leibler divergence for generic graphs using our method has a remarkably simple form, i.e., a symmetric function of eigenvalues of concentration matrices. This fact establishes the connection of structure recovery limits and eigenvalues of concentration matrices. Essentially, the extreme simplicity of final results stem from the inherent symmetries of orthogonal group. And we believe that inherent symmetries play a crucial role in graphical models learning, which is worthy of further investigation. Last but not the least, while we have considered only GMRF structure learning in this paper, our analysis framework can be extended to other structure learning problems of undirected graphical models.

## 2 GMRF Structure Learning

We begin with some background on Gaussian Markov random fields and graphical model selection problem. Given an undirected graph $G = (V, E)$, with a collection $V = \{1, \dots, p\}$ of vertices joined by a collection $E \subseteq V \times V$ of undirected edges. A Gaussian random field is obtained by associate a scalar Gaussian random variable $X_i$ with each vertex $i$, and then specifying a joint Gaussian distribution over the random vector $X = (X_1, \dots, X_p)$. In this paper, $X$ has zero mean and covariance matrix $\Sigma$ which is assumed to be positive-definite. Accordingly, its probability density function for $x \in \mathbb{R}^p$ has the form

$$\varphi(x; \Theta) = \frac{1}{\sqrt{(2\pi)^p \det(\Theta)^{-1}}} \exp\left\{-\frac{1}{2} x' \Theta x\right\},$$

where $\Theta = \Sigma^{-1}$ is the inverse covariance or concentration matrix and we use the notation $x' \equiv x$ transposed. For a Gaussian Markov random field, in addition to a Gaussian random field, the non-zero structure of $\Theta$ is specified by the associated graph structure. More precisely, $\Theta_{ij} = 0$ if $(i, j) \notin E$ by the Hammersley-Clifford theorem (Lauritzen, 1996).

The task of graphical model selection is to estimate the underlying graph $G$ based on the $n$ i.i.d. observations $X_1^n := \{X_1, \dots, X_n\}$, which is highly dependent of values of the inverse covariance matrix entries. In this paper, we consider the collection of $\mathcal{G}_{p,k}(\lambda)$ of Gaussian Markov random fields with $p$ vertices, $k$ edges and a positive lower bound $\lambda \in (0, 1)$ on the minimum value of non-zero off-diagonal matrix elements in $\Theta$. A decoder $\phi : \mathbb{R}^{n\times p} \to \mathcal{G}_{p,k}(\lambda)$ is a mapping from $X_1^n$ to an estimated graph. For any decoder $\phi$, the maximal error probability over the family $\mathcal{G}_{p,k}(\lambda)$ is defined as

$$q_{\max}(\phi) = \max_{G \in \mathcal{G}_{p,k}(\lambda)} \mathbb{P}\{\phi(X_1^n) \neq G\}.$$

We use the following standard asymptotics notations. $a_n = \mathcal{O}(b_n)$ means that $a_n \leqslant C_1 b_n$ for some constant $C_1 > 0$ and $a_n = \Omega(b_n)$ means that $a_n \geqslant C_2 b_n$ for some constants $C_2 > 0$. And $a_n = \Theta(b_n)$ is shorthand for $a_n = \mathcal{O}(b_n)$ and $a_n = \Theta(b_n)$.

## 3 Main Results and Consequences

In this section, we present main results of the structure learning limit. The analysis and proof are deferred to the next section.

**Theorem 3.1.** *Given $n$ i.i.d. observations $X_1^n := \{X_1, \dots, X_n\}$, then a necessary condition for asymptotically reliable recovery over $\mathcal{G}_{p,k}(\lambda)$ is*

$$n > \frac{p\left(\log\left[\binom{p}{2} - k + 1\right] - 1\right)}{4k\lambda}.$$

Firstly, we notice that $\lambda$ has a direct effect on the difficulty of graphical model selection from the theorem 3.1, i.e., more samples are required for small values of $\lambda$. Furthermore, we explore recovery limits under two scalings of edge sparsity: the regime of linear edge sparsity ($k = \Theta\left(\binom{p}{2}\right)$) and sublinear edge sparsity ($k = o\left(\binom{p}{2}\right)$), as shown in Table 1. In each regime, we consider two important kinds of scalings of $\lambda = \Theta\left(\frac{1}{k}\right)$ and $\lambda = \Theta(1)$. It's important to realize that the result in the theorem 3.1 roots in eigenvalues analysis of concentration matrices in an ensemble to a certain extent, and it could be improved by making a more thorough investigation of eigenvalue properties.

Table 1: The Necessary Conditions on the Number of Samples Required for Exact Graph Recovery

| PARAMETER REGIMES | LOWER BOUNDS |
|---|---|
| $k = \Theta\left(\binom{p}{2}\right)$ $\lambda = \Theta\left(\frac{1}{k}\right)$ | $\Theta\left(p \log \binom{p}{2}\right)$ |
| $k = \Theta\left(\binom{p}{2}\right)$ $\lambda = \Theta(1)$ | $\Theta\left(\frac{p}{k} \log \binom{p}{2}\right)$ |
| $k = o\left(\binom{p}{2}\right)$ $\lambda = \Theta\left(\frac{1}{k}\right)$ | $\Theta\left(p \log \left[\binom{p}{2} - k\right]\right)$ |
| $k = o\left(\binom{p}{2}\right)$ $\lambda = \Theta(1)$ | $\Theta\left(\frac{p}{k} \log \left[\binom{p}{2} - k\right]\right)$ |

It is worthwhile comparing our result to existing necessary and sufficient conditions on model selection for Gaussian Markov random fields. We note that Wang et al. (Wang et al., 2010) have shown that a necessary condition provided by for consistent graph selection for $\mathcal{G}_{p,d}(\lambda)$ with $\lambda \in [0, \frac{1}{2}]$ is $n > \max\{c_1 d^2 \log(p - d), c_2 d^{1-\epsilon} \log(\frac{p}{d})\}$ for some constants $c_1, c_2 > 0$ and any $\epsilon > 0$ in the regime of $\lambda = \Theta(\frac{1}{d})$, where $\mathcal{G}_{p,d}(\lambda)$ is a family of graphs on $p$ nodes with edge sets that have degree at most $d$. Since any model in $\mathcal{G}_{p,k}(\lambda)$ has degree at most $k$, it follows from (Wang et al., 2010) that the necessary sample size for exact reliable graphical model selection for $\mathcal{G}_{p,k}(\lambda)$ will be more than $\max\{c_1 k^2 \log(p - k), c_2 k^{1-\epsilon} \log(\frac{p}{k})\}$. Thereby, in the same regime of $\lambda = \Theta(\frac{1}{d})$, our result becomes $\Theta\left(\frac{pd}{k} \log \left[\binom{p}{2} - k\right]\right)$ which provides a sharper limit than the result in (Wang et al., 2010) in certain situations such as sparse graphical models.

For sufficient conditions on model selection for Gaussian Markov random fields, the $\ell_1$-regularized Gaussian MLE (Graphical Lasso) methods (d' Aspremont et al., 2008)(Friedman et al., 2008)(Ravikumar et al., 2008)(Yuan and Yin, 2007) require $\Omega(d^2 \log p)$ samples with high probability under the irrepresentable condition. And the nodewise $\ell_1$-regularization linear regression (Neighborhood Lasso) (Meinshausen and B$\ddot{u}$hlmann, 2006) requires $\Omega(d \log p)$ samples (Meinshausen and B$\ddot{u}$hlmann, 2006)(Wainwright, 2009) to guarantee sparsistency using some regularity assumptions for the covariance matrices and neighborhood stability assumption. More recently,"stagewise" greedy methods (Jalali et al., 2011)(Johnson et al., 2012) require $\Omega(d \log p)$ samples for sparsistent graph recovery under a restricted eigenvalue and restricted smoothness condition on the true concentration matrix. It can be seen that some existing sufficient conditions require less samples than our necessary result, which is due to various conditions imposed upon estimators. In general, a consistent graph selection may require much more samples without any assumption on the true concentration matrix.

## 4 Analysis: From Symmetry to Limit

This section is devoted to our formal analysis of structure learning limits. Suppose that the decoder is told the locations of all but the smallest non-zero off-diagonal value of $\Theta$, as well as the values of $\Theta$ on its off-diagonal support. The remaining sub-problem is to determine, given the $n$ observations $X_1^n := \{X_1, \ldots, X_n\}$, the location of the smallest nonzero value of $\Theta$. It is clear that the error probability of the decoder in this restricted problem provides a lower bound on the error probability in the original problem.

### 4.1 Starting with Fano

We perform our analysis with applying Fano's method to the whole ensemble $\mathcal{E}$, which consists of $\binom{p}{2} - k + 1$ graphs that contain the given $k - 1$ edges in $\mathcal{G}_{p,k}(\lambda)$. Suppose a graph index $\Xi$ is chosen uniformly over $\{1, \ldots, |\mathcal{E}|\}$. Given $n$ i.i.d. samples $X_1^n$, by Fano's inequality (Cover and Thomas, 1991), the maximal error probability is lower bounded as

$$q_{\max}(\phi) \geqslant 1 - \frac{I(\Xi; X_1^n) + 1}{\log |\mathcal{E}|}, \qquad (1)$$

where $I(\Xi; X_1^n)$ denotes the mutual information of $\Xi$ and $X_1^n$. We take log to base 2 throughout this paper. Consequently, the problem is reduced to analyze the mutual information $I(\Xi; X_1^n)$. Let $G_S$ and $G_T$ be a pair of distinct graphs in $\mathcal{E}$ with corresponding concentration matrices $S$ and $T$, then it is easy to obtain the following upper bound of the mutual information by convexity of the Kullback-Leibler divergence

$$I(\Xi; X_1^n) \leqslant \frac{n}{|\mathcal{E}|^2} \sum_{(G_S, G_T) \in \mathcal{E} \times \mathcal{E}} \mathbb{S}(\mathbb{P}_{x|S} \| \mathbb{P}_{x|T}), \qquad (2)$$

where $\mathbb{S}(\mathbb{P}_{x|S} \| \mathbb{P}_{x|T})$ is the symmetric Kullback-Leibler divergence between the distributions $\mathbb{P}_{x|S}$ and $\mathbb{P}_{x|T}$, defined in the natural way via

$$\mathbb{S}(\mathbb{P}_{x|S} \| \mathbb{P}_{x|T}) = D(\mathbb{P}_{x|S} \| \mathbb{P}_{x|T}) + D(\mathbb{P}_{x|T} \| \mathbb{P}_{x|S}).$$

### 4.2 The Geometry of Concentration Matrices

This section presents a crucial geometry property of concentration matrices that they can be transformed into each other by a space reflection or rotation. Throughout the paper, we take the notation of

$\Omega^{(\alpha\beta)} = I_{|V|\times|V|} - E_{\alpha\alpha} - E_{\beta\beta} + E_{\alpha\beta} + E_{\beta\alpha}$, where $E_{\alpha\beta}$ is $|V|\times|V|$ matrix with the $(\alpha,\beta)$ entry equal to 1 and 0 entries elsewhere. For $G_\Theta \in \mathcal{E}$, let $\Theta_{ii} = 1+\mu > 1$ if the node $i$ has a positive degree, otherwise $\Theta_{ii} = 1$. One of the crucial properties used in the subsequent section will be the following orthogonal relationship of concentration matrices

**Lemma 4.1.** *For any two concentration matrices $S$ and $T$, there exists an orthogonal matrix $\Omega$ satisfying $T = \Omega S \Omega$.*

**Proof**

It is not difficult to verify that there are two cases for different $G_S, G_T \in \mathcal{E}$:

$$\text{(Case I)}\quad \begin{array}{l} E(G_S)\backslash E(G_T) = \{(\gamma,\alpha)\} \quad \text{and} \\ E(G_T)\backslash E(G_S) = \{(\gamma,\beta)\}, \end{array}$$

$$\text{(Case II)}\quad \begin{array}{l} E(G_S)\backslash E(G_T) = \{(\alpha,\beta)\} \quad \text{and} \\ E(G_T)\backslash E(G_S) = \{(\gamma,\delta)\}, \end{array}$$

where $\alpha,\beta,\gamma,\delta \in V$ and $E(G_S)$ and $E(G_T)$ are edges sets of $G_S$ and $G_T$. Then in the case (I), there is an orthogonal matrix $\Omega^{(\alpha\beta)}$ such that

$$T = \Omega^{(\alpha\beta)} S \Omega^{(\alpha\beta)}, \tag{3}$$

where $\Omega^{(\alpha\beta)}$ is an improper orthogonal matrix since $\det \Omega^{(\alpha\beta)} = -1$. Equivalently saying, $S$ can be transformed into $T$ by the interchange of the $\alpha$-th row and $\beta$-th row of $S$ first and then the interchange of the $\alpha$-th column and $\beta$-th column of $S$. In the case (II), it is easy to check that the matrices $S$ and $T$ must satisfy the relations

$$T = \Omega^{(\alpha\delta)}\Omega^{(\beta\gamma)} S \Omega^{(\beta\gamma)}\Omega^{(\alpha\delta)} \tag{4}$$

for two improper orthogonal matrices $\Omega^{(\alpha\delta)}$ and $\Omega^{(\beta\gamma)}$. Hence, the lemma follows from Eq. (3) and Eq. (4). □

We take as an example as an illustration of Lemma 4.1. Suppose $V = \{1,2,3,4\}$ and $E(G_S) = \{(1,2),(1,3)\}, E(G_T) = \{(1,3),(2,3)\}$, $\mu,\theta,\tau > 0$, then $T = \Omega^{(12)} S \Omega^{(12)}$ follows from the following equality

$$\begin{pmatrix} 1+\mu & & \tau & \\ & 1 & \theta & \\ \tau & \theta & 1+\mu & \\ & & & 1 \end{pmatrix} = \begin{pmatrix} 0 & & 1 & \\ & 1 & & \\ 1 & & 0 & \\ & & & 1 \end{pmatrix}$$
$$\begin{pmatrix} 1+\mu & \theta & \tau & \\ \theta & 1 & & \\ \tau & & 1+\mu & \\ & & & 1 \end{pmatrix} \begin{pmatrix} 0 & & 1 & \\ & 1 & & \\ 1 & & 0 & \\ & & & 1 \end{pmatrix}.$$

From the geometric perspective, two concentration matrices are related by a space reflection or rotation in Euclidean space. More specifically, $\Omega$ induces a reflection of $\mathbb{R}^p$ across a axis in the case I. In the case II, $\Omega$ induces twice reflections of $\mathbb{R}^p$ across two axes, or equivalently, a rotation of $\mathbb{R}^p$. Refer to (Cartan, 1928)(Weyl, 1939) for more background information.

### 4.3 Eigensystems of Concentration Matrices

Since $S$ is a real symmetric matrix, there is an orthogonal matrix $\Gamma$ such that

$$S = \Gamma'\Lambda\Gamma = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_p \end{pmatrix}' \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix} \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_p \end{pmatrix},$$

where $\{\lambda_k, \xi_k\}_{k=1}^p$ are eigenvalues and corresponding row eigenvectors of $S$ with unit modulus.

Recall that there is an orthogonal matrix $\Omega$ satisfying $T = \Omega S \Omega$. Let

$$\Gamma\Omega = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_p \end{pmatrix} \Omega = \begin{pmatrix} \xi_1\Omega \\ \vdots \\ \xi_p\Omega \end{pmatrix} = \begin{pmatrix} \xi_1^* \\ \vdots \\ \xi_p^* \end{pmatrix},$$

then we claim that $\{\lambda_k, \xi_k^*\}_{k=1}^p$ are eigenvalues and corresponding eigenvectors of $T$. This can be seen from noticing that $\xi_k\Omega = \xi_k^*$ implies $\xi_k^*\Omega = \xi_k\Omega^2 = \xi_k$, which lead to the following simple calculation

$$\xi_k^* T = \xi_k^*\Omega S \Omega = \xi_k S \Omega = \lambda_k \xi_k \Omega = \lambda_k \xi_k^*.$$

From the geometric perspective, $\{\xi_k\}_{k=1}^p$ and $\{\xi_k^*\}_{k=1}^p$ are two orthonormal frames of unit vectors which always differ by a space reflection or rotation. And we can call $\{\lambda_k, \xi_k\}_{k=1}^p$ and $\{\lambda_k, \xi_k^*\}_{k=1}^p$ as corresponding eigensystems of $S$ and $T$.

### 4.4 The Orthogonal Invariance of the Symmetric Kullback-Leibler Divergence

In this section, we investigate the invariance of orthogonal group actions on the symmetric Kullback-Leibler divergence. More explicitly, the crucial property of the symmetric Kullback-Leibler divergence $\mathbb{S}(\mathbb{P}_{x|S}\|\mathbb{P}_{x|T})$ is the invariance under orthogonal transforms of $S$ and $T$. We start with simplifying the form of the symmetric Kullback-Leibler divergence as stated in the following lemma.

**Lemma 4.2.** *For any two concentration matrices $S$ and $T$, we have the symmetric Kullback-Leibler divergence $\mathbb{S}(\mathbb{P}_{x|S}\|\mathbb{P}_{x|T})$ between the distributions $\mathbb{P}_{x|S}$ and $\mathbb{P}_{x|T}$ as follows*

$$\frac{1}{\sqrt{(2\pi)^p \det(S)^{-1}}} \int_{\mathbb{R}^p} (x'Tx - x'Sx) \exp\left(-\frac{1}{2}x'Sx\right) dx.$$

**Proof**

It is straightforward to show that

$$\begin{aligned}&\mathbb{S}(\mathbb{P}_{x|S}\|\mathbb{P}_{x|T})\\ =&\int_{\mathbb{R}^p}\varphi(x;S)\ln\frac{\varphi(x;S)}{\varphi(x;T)}dx+\int_{\mathbb{R}^p}\varphi(x;T)\ln\frac{\varphi(x;T)}{\varphi(x;S)}dx\\ =&\frac{1}{2\sqrt{(2\pi)^p\det(S)^{-1}}}\Big[\int_{\mathbb{R}^p}(x'Tx-x'Sx)\exp\left(-\frac{1}{2}x'Sx\right)dx\\ &+\int_{\mathbb{R}^p}(x'Sx-x'Tx)\exp\left(-\frac{1}{2}x'Tx\right)dx\Big]\end{aligned}\tag{5}$$

where we apply the fact of $\det(S)=\det(T)$ since $S$ and $T$ have the same eigenvalues. With the aid of Lemma 4.1, there is an orthogonal matrix $\Omega$ satisfying $T=\Omega S\Omega$. Then, write $x=\Omega y$ yields

$$\begin{aligned}&\int_{\mathbb{R}^p}(x'Sx-x'Tx)\exp\left(-\frac{1}{2}x'Tx\right)dx\\ =&|\det(\Omega)|\int_{\mathbb{R}^p}[(\Omega y)'S(\Omega y)-(\Omega y)'T(\Omega y)]\\ &\exp\left(-\frac{1}{2}(\Omega y)'T(\Omega y)\right)dy\\ =&\int_{\mathbb{R}^p}(y'Ty-y'Sy)\exp\left(-\frac{1}{2}y'Sy\right)dy\\ =&\int_{\mathbb{R}^p}(x'Tx-x'Sx)\exp\left(-\frac{1}{2}x'Sx\right)dx.\end{aligned}\tag{6}$$

We complete the proof of Lemma 4.2 by combing Eq. (5) and Eq. (6). □

We proceed to prove the orthogonal invariance of the symmetric Kullback-Leibler divergence, which is crucial for our analysis.

**Lemma 4.3.** *For any two concentration matrices $S$ and $T$, the symmetric Kullback-Leibler divergence $\mathbb{S}(\mathbb{P}_{x|S}\|\mathbb{P}_{x|T})$ is invariant under transformations of $S\to\Omega'S\Omega$ and $T\to\Omega'T\Omega$ for an orthogonal matrix $\Omega$.*

**Proof**

$$\begin{aligned}&\mathbb{S}(\mathbb{P}_{x|\Omega'S\Omega}\|\mathbb{P}_{x|\Omega'T\Omega})\\ =&\frac{1}{\sqrt{(2\pi)^p\det(\Omega'S\Omega)^{-1}}}\int_{\mathbb{R}^p}(x'\Omega'T\Omega x-x'\Omega'S\Omega x))\\ &\exp\left(-\frac{1}{2}x'\Omega'S\Omega x\right)dx\\ =&\frac{1}{\sqrt{(2\pi)^p\det(S)^{-1}}}\int_{\mathbb{R}^p}[(\Omega x)'T(\Omega x)-(\Omega x)'S(\Omega x)]\\ &\exp\left(-\frac{1}{2}(\Omega x)'S(\Omega x)\right)dx.\end{aligned}$$

Changing the variable of integration by letting $y=\Omega x$, we get

$$\begin{aligned}&\mathbb{S}(\mathbb{P}_{x|\Omega'S\Omega}\|\mathbb{P}_{x|\Omega'T\Omega})\\ =&\frac{1}{|\det(\Omega)|\sqrt{(2\pi)^p\det(S)^{-1}}}\\ &\int_{\mathbb{R}^p}(y'Ty-y'Sy)\exp\left(-\frac{1}{2}y'Sy\right)dy\\ =&\frac{1}{\sqrt{(2\pi)^p\det(S)^{-1}}}\int_{\mathbb{R}^p}(x'Tx-x'Sx)\exp\left(-\frac{1}{2}x'Sx\right)dx\\ =&\mathbb{S}(\mathbb{P}_{x|S}\|\mathbb{P}_{x|T}).\end{aligned}$$

□

### 4.5 The KL Divergence as a Symmetric Function of Eigenvalues

We are in a position to analyze the symmetric Kullback-Leibler divergence $\mathbb{S}(\mathbb{P}_{x|S}\|\mathbb{P}_{x|T})$ by applying lemmas developed in section 4.4. From Lemma 4.2, we see that $x'Tx-x'Sx$ is the key factor in analyzing $\mathbb{S}(\mathbb{P}_{x|S}\|\mathbb{P}_{x|T})$. Therefore, our first purpose is to show that the expression of $x'Tx-x'Sx$ can be expressed using eigenvalues and eigenvectors of $S$ and $T$, as stated in the following lemma.

**Lemma 4.4.** *For any two concentration matrices $S$ and $T$, we have*

$$x'Tx-x'Sx=\sum_{k=1}^{p}\lambda_k(\xi_k^*+\xi_k)x(\xi_k^*-\xi_k)x,$$

*where $\{\lambda_k,\xi_k\}_{k=1}^p$ are eigenvalues and eigenvectors of $S$, while $\{\lambda_k,\xi_k^*\}_{k=1}^p$ are eigenvalues and eigenvectors of $T$.*

**Proof**

$$\Gamma x=\begin{pmatrix}\xi_1\\ \vdots\\ \xi_p\end{pmatrix}x=\begin{pmatrix}\xi_1 x\\ \vdots\\ \xi_p x\end{pmatrix}$$

gives

$$x'Sx=x'\Gamma'\Lambda\Gamma x=(\Gamma x)'\Lambda(\Gamma x)=\sum_{k=1}^{p}\lambda_k(\xi_k x)^2.$$

Since

$$\Gamma\Omega x=\begin{pmatrix}\xi_1^*\\ \vdots\\ \xi_p^*\end{pmatrix}x=\begin{pmatrix}\xi_1^* x\\ \vdots\\ \xi_p^* x\end{pmatrix},$$

we obtain

$$x'Tx=x'\Omega S\Omega x=x'\Omega\Gamma'\Lambda\Gamma\Omega x$$

and thus $x'Tx = (\Gamma\Omega x)'\Lambda(\Gamma\Omega x) = \sum_{k=1}^{p} \lambda_k(\xi_k^* x)^2$. Hence $x'Tx - x'Sx$ has the following expression

$$\sum_{k=1}^{p} \lambda_k[(\xi_k^* x)^2 - (\xi_k x)^2] = \sum_{k=1}^{p} \lambda_k(\xi_k^* + \xi_k)x(\xi_k^* - \xi_k)x.$$

□

With the help of the previous Lemma 4.4, we can establish the relationship between the symmetric Kullback-Leibler divergence and eigenvalues and concentration matrices.

**Theorem 4.1.** *For any two concentration matrices $S$ and $T = \Omega S \Omega$ for certain orthogonal matrix $\Omega$, then the symmetric Kullback-Leibler divergence $\mathbb{S}(\mathbb{P}_{x|S}\|\mathbb{P}_{x|T})$ has the form*

$$\mathbb{S}(\mathbb{P}_{x|S}\|\mathbb{P}_{x|T}) = \begin{cases} \frac{\lambda_\alpha}{\lambda_\beta} + \frac{\lambda_\beta}{\lambda_\alpha} - 2 & \textit{Case I} \\ \\ \frac{\lambda_\alpha}{\lambda_\delta} + \frac{\lambda_\delta}{\lambda_\alpha} + \frac{\lambda_\beta}{\lambda_\gamma} + \frac{\lambda_\gamma}{\lambda_\beta} - 4 & \textit{Case II,} \end{cases}$$

*where $\Omega = \Omega^{(\alpha\beta)}$ in the case I and $\Omega = \Omega^{(\alpha\delta)}\Omega^{(\beta\gamma)}$ in the case II. $\lambda_\alpha, \lambda_\beta$ are eigenvalues of $S$ in the case I and $\lambda_\alpha, \lambda_\beta, \lambda_\delta, \lambda_\gamma$ are eigenvalues of $S$ in the case II.*

The proof of this theorem is applying lemmas developed in section 4.4 and Lemma 4.4. The proof involves straightforward but length calculations and will be given in the appendix. It is worth noting that final result of the symmetric Kullback-Leibler divergence is extremely simple, which is a symmetric function of eigenvalues of concentration matrices. We believe that it is the inherent symmetry of orthogonal systems that leads to this remarkable final result.

## 4.6 Bounding the Symmetric Function of Eigenvalues

It can be seen from 4.1 that eigenvalues of concentration matrices play a crucial role in measuring differences between models. To analyze the connection between the average symmetric Kullback-Leibler divergence and eigenvalues of concentration matrices, we introduce an useful symmetric function of $(\lambda_\alpha, \lambda_\beta)$ as $\psi_{\alpha\beta} = \lambda_\alpha\lambda_\beta^{-1} + \lambda_\beta\lambda_\alpha^{-1} - 2$. The next theorem deals with the upper bound of the average value of $\psi_{\alpha\beta}$, which will lead to the final information-theoretic limit.

**Theorem 4.2.** *For any two concentration matrices $S$ and $T$ corresponding two models in $\mathcal{G}_{p,k}(\lambda)$, let $\psi_{\alpha\beta} = \lambda_\alpha\lambda_\beta^{-1} + \lambda_\beta\lambda_\alpha^{-1} - 2$, then the average value of $\psi_{\alpha\beta}$ is upper bounded by*

$$\frac{1}{p^2}\sum_{\alpha,\beta=1}^{p} \psi_{\alpha\beta} < \frac{2k\lambda}{p},$$

*where $\lambda_\alpha$ and $\lambda_\beta$ are defined in the same meaning in Theorem 4.1.*

### Proof

It is easy to verify that the worse case error probability occurs if $S_{ij} = \lambda$ for $(i,j) \in E(G_S)$. Evidently one has $S$ is unit matrix in the initial state of $k = 0$. As $k$ increases by one, two off-diagonal elements $S_{ij}$ and $S_{ji}$ are set to a positive number $\lambda$, and it is easy to verify that $\sum_{\mu=1}^{p} S_{\mu\mu}$ increases at most $2\lambda$, yielding the fact that $\sum_{\mu=1}^{p} \lambda_\mu = \sum_{\mu=1}^{p} S_{\mu\mu} \leqslant p + 2k\lambda$.

If $\sum_{\mu=1}^{p} \lambda_\mu < p + 2k\lambda$, assume $\lambda_1$ is the largest value of $\{\lambda_1, \cdots, \lambda_p\}$, then setting $\lambda_1' = \lambda_1 + (p + 2k\lambda - \sum_{\mu=1}^{p} \lambda_\mu)$ leads to a larger value of $\sum_{\alpha,\beta=1}^{p} \psi_{\alpha\beta}$. So $\sum_{\alpha,\beta=1}^{p} \psi_{\alpha\beta}$ attains the maximum when $\sum_{\mu=1}^{p} \lambda_\mu = p + 2k\lambda$. Furthermore, $\sum_{\mu=1}^{p} \lambda_\mu^{-1} \leqslant p - 1 + (2k\lambda + 1)^{-1}$ since if $\lambda_\alpha > 1$ and $\lambda_\beta > 1$, we can set $\lambda_\alpha' = 1$ and $\lambda_\beta' = \lambda_\alpha + \lambda_\beta - 1$ which remains $\sum_{\mu=1}^{p} \lambda_\mu$ unchanged and makes $\sum_{\mu=1}^{p} \lambda_\mu^{-1}$ larger. So $\sum_{\mu=1}^{p} \lambda_\mu^{-1}$ attains the maximum when $p - 1$ values of $\lambda_\mu$ are equal to 1.

Finally, notice that $\psi_{\alpha\beta} \leqslant \lambda_\beta + \lambda_\beta^{-1} - 2$ if $\lambda_\alpha \geqslant \lambda_\beta$. Re-ordering indices as $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$, we obtain the following estimate

$$\begin{aligned} &\sum_{\alpha,\beta=1}^{p} \psi_{\alpha\beta} \\ \leqslant &\sum_{\mu=1}^{p-1} \left[ (\mu - 1)\left(\lambda_\mu + \frac{1}{\lambda_\mu} - 2\right) + \sum_{\nu=\mu+1}^{p} \left(\lambda_\nu + \frac{1}{\lambda_\nu} - 2\right) \right] \\ &+ (p-1)\left(\lambda_p + \frac{1}{\lambda_p} - 2\right) \\ =&2\left[\sum_{\mu=2}^{p}(\mu-1)\lambda_\mu + \sum_{\mu=2}^{p}\frac{\mu-1}{\lambda_\mu} - p(p-1)\right] \\ \leqslant&2\Bigg[\frac{p-1}{2}(p + 2k\lambda) + \frac{p-1}{2}\left(p - 1 + \frac{1}{2k\lambda+1}\right) \\ &- p(p-1)\Bigg] \\ =&(p-1)\left(2k\lambda + \frac{1}{2k\lambda+1} - 1\right). \end{aligned}$$

Hence, the average value of $\psi_{\alpha\beta}$ is upper bounded by

$$\begin{aligned} &\frac{1}{p^2}\sum_{\alpha,\beta=1}^{p} \psi_{\alpha\beta} \\ \leqslant&\frac{p-1}{p^2}\left(2k\lambda + \frac{1}{2k\lambda+1} - 1\right) \\ <&\frac{1}{p}\left(2k\lambda + \frac{1}{2k\lambda+1} - 1\right) \\ <&\frac{2k\lambda}{p}. \end{aligned}$$

□

### 4.7 Final Information-Theoretic Limit

We are ready to proceed with the final step, which will prove the main theorem in this paper, i.e., Theorem 3.1.

It is evident the average value of $\psi_{\alpha\beta}$ in the remaining possible models keeps non-increasing as non-zero off-diagonal elements of the concentration matrix are determined in descending order.

Applying Theorem 4.1 and Theorem 4.2 immediately gives

$$\frac{1}{|\mathcal{E}|^2} \sum_{(S,T)\in\mathcal{E}\times\mathcal{E}} \mathbb{S}(\mathbb{P}_{x|S}\|\mathbb{P}_{x|T}) \leqslant 2 \cdot \frac{1}{p^2} \sum_{\alpha,\beta=1}^{p} \psi_{\alpha\beta} < \frac{4k\lambda}{p}. \tag{7}$$

Applying inequalities (1), (2) and (7) with $|\mathcal{E}| = \binom{p}{2} - k + 1$

$$q_{\max}(\phi) > 1 - \frac{1}{\log\left[\binom{p}{2} - k + 1\right]} \left(\frac{4k\lambda n}{p} + 1\right).$$

Consequently, a necessary condition for asymptotically reliable recovery over $\mathcal{G}_{p,k}(\lambda)$ is

$$n > \frac{p\left(\log\left[\binom{p}{2} - k + 1\right] - 1\right)}{4k\lambda}.$$

Thus, the main theorem in this paper is proved.

## 5 Concluding Remarks

In this paper, we analyze the information-theoretic limit on consistent model selection for Gaussian Markov random fields. We treat this problem from a group theoretical viewpoint, i.e., the invariance under orthogonal group actions on concentration matrices and the symmetric Kullback-Leibler divergence. Our analysis reveals the connection between the graphical model selection limit and eigenvalues of concentration matrices.

It is worth noting that the symmetric Kullback-Leibler divergence has a rather simple and symmetric expression in terms of of eigenvalues of concentration matrices. We believe that it is inherent symmetries of concentration matrices and the symmetric Kullback-Leibler divergence that lead to final remarkable expression. And the inherent symmetries in graphical models merit further investigation. Last but not the least, while we have considered only GMRF structure learning in this paper, our analysis framework can be extended to other structure learning of undirected graphical models.

**Acknowledgements**

We thank anonymous reviewers for helpful comments. This research is supported by 973 Program (2013CB329503)NSFC (Grant No. 91120301) and Beijing Municipal Education Commission Science and Technology Development Plan key project under grant KZ201210005007.

### References

A. Ahmedy, L. Song, and E. P. Xing. Time-varying networks: recovering temporally rewiring genetic networks during the life cycle of drosophila melanogaster. arXiv:0901.0138. 2008.

A. Anadkumar, V. Y. F. Tan, and A. S. Wilsky. High-dimensional structure learning of Ising Models on sparse random graphs. arXiv:1011.0129. 2010.

A. Anadkumar, V. Y. F. Tan, and A. S. Wilsky. High-dimensional Gaussian graphical model selection: tractable graph families. arXiv:1107.1270v2. 2011.

A. d'Aspremont, O. Banerjee, and L. E. Ghaoui. The first order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications*, **30**(1):56-66, 2008.

G. Bresler, E. Mossel, A. Sly. Reconstruction of Markov random fields from samples: some observations and algorithms. In: *International Workshop APPROX Approximation, Randomization and Combinatorial Optimization*, pp. 343-356, Springer, Heidelberg, 2008.

É. Cartan. *Riemannian Geometry in an Orthogonal Frame: From Lectures Delivered by Élie Cartan at the Sorbonne in 1926-1927*, 1928.

T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley and Sons, New York, 1991.

R. Durbin, S. Eddy, A. Krogh, and G. Mitchison editors. *Biological Sequnence Analysis.* Cambridge University Press, Cambridge, 1998.

S. Geman and D. Geman. Stochatic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. PAMI*, **6**:721-741, 1984.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3):432-441, 2007.

C. J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar. Sparse inverse covariance matrix estimation using quadratic approximation. *Advances in Neural Information Processing Systems*, 2011.

A. Jalali, C. C. Johnson, and P. Ravikumar. On learning discrete graphical models using greedy methods.

*Advances in Neural Information Processing Systems*, 2011.

C. C. Johnson, A. Jalali, and P. Ravikumar. High-dimensional sparse inverse covariance estimation using greedy methods. *International Conference on Artificial Intelligence and Statistics*, 2012.

S. L. Lauritzen. *Graphical Model.* Oxford University Press, Oxford, 1996.

H. Weyl. *The Classical Groups.* Princeton University Press, Princeton, N. J. (1939)

N. Meinshausen and P. B*ü*hlmann. High-dimensional graphs and variable selection with the lasso. *Annuals of Statistics*, **34**:1436-1462, 2006.

P. Ravikumar, M. J. Wainwright, and L. Lafferty. High-dimensional Ising model selection using $\ell_1$-regularized logistic regression. *Annals of Statistics*, **38**(3):1287-1319, 2008.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Technical report*, Department of Statistics, UC Berkeley, 2008.

A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invarianct covariance estimation. *Electronic Journal of Statistics*, **2**:494-515, 2008.

N. Santhanam and M. J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. In *International Symposium on Information Theory*, IEEE Press, Toronto, Canada, 2008.

V. R. Fernando. *Complex Social Networks* (Econometric Society Monographs). Cambridge University Press, Cambridge, 2007.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (lasso). *IEEE Trans. Info. Theory*, **55**:2183-2202, 2009.

W. Wang, M. J. Wainwright, K. Ramchandran. Information-theoretic bounds on model selection for Gaussian Markov random fields. In *IEEE International Symposium on Information Theory*, IEEE Press, Austin, TX, 2010.

S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications.* Cambridge University Press, New York, NY, 1994.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergences. *Annals of Statistics* **27**:1564-1599, 1999.

B. Yu. Assouad, Fano and Le Cam. Research papers in probability and statistics: Festschrift in Hanor of Lucien Le Cam 423-435, 1996.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**:19-35, 2007.

# Appendix

## 5.1 Proof of Theorem 4.1

Since $S$ is a real symmetric matrix, there is an orthogonal matrix $\Gamma$ such that $S = \Gamma'\Lambda\Gamma$, where $\Gamma = \{\lambda_k\}_{k=1}^p$ and $\{\xi_k\}_{k=1}^p$ are eigenvalues and corresponding row eigenvectors of $S$ with unit modulus. Then an orthogonal transformation of $S$ leads to an orthogonal transformation of $\Gamma$. Therefore, the orthogonal matrix $\Gamma$ can be transformed into $\mathrm{diag}(\pm 1, \cdots, \pm 1)$ by performing certain orthogonal transform of $S \to \Omega' S \Omega$. Correspondingly, $T$ is transformed by $T \to \Omega' T \Omega$. From Lemma 4.3, we see that $\mathbb{S}(\mathbb{P}_{x|S} \| \mathbb{P}_{x|T}) = \mathbb{S}(\mathbb{P}_{x|\Omega' S \Omega} \| \mathbb{P}_{x|\Omega' T \Omega})$. Henceforth, we may assume that elements of $\xi_k$ is all zero but the $k^{th}$ element is $\pm 1$. In the case I, $\xi_k$ is transformed to $\xi_k^*$ by means of an improper orthogonal transformation $\Omega^{(\alpha\beta)}$.

Write $x = (x_1, \ldots, x_p)'$ and $\xi_k = (\xi_{k1}, \ldots, \xi_{kp})$. By reordering indices as needed be, we may assume without loss of generality that $1 < \alpha < \beta < p$, then $\xi_k^* = \xi_k \Omega^{(\alpha\beta)}$ gives rise to

$$\xi_k^* + \xi_k = (2\xi_{k1}, \cdots, \xi_{k\alpha} + \xi_{k\beta}, \cdots, \xi_{k\alpha} + \xi_{k\beta}, \cdots, 2\xi_{kp}),$$
$$\xi_k^* - \xi_k = (0, \cdots, \xi_{k\beta} - \xi_{k\alpha}, \cdots, \xi_{k\alpha} - \xi_{k\beta}, \cdots, 0),$$

where the intermediate terms are the $\alpha^{th}$ and $\beta^{th}$ positions of $\xi_k^* + \xi_k$ and $\xi_k^* - \xi_k$, respectively. Thereby,

$$\begin{aligned}&(\xi_k^* + \xi_k)x(\xi_k^* - \xi_k)x \\ =&2\xi_k x(\xi_{k\beta} - \xi_{k\alpha})(x_\alpha - x_\beta) + [(\xi_{k\beta} - \xi_{k\alpha})(x_\alpha - x_\beta)]^2.\end{aligned}$$

Applying Lemma 4.4, we obtain

$$\begin{aligned}&\int_{\mathbb{R}^p} (x'Tx - x'Sx) \exp\left(-\frac{1}{2}x'Sx\right) dx \\ =&2\sum_{k=1}^p \lambda_k(\xi_{k\beta} - \xi_{k\alpha}) \int_{\mathbb{R}^p} \xi_k x(x_\alpha - x_\beta) \exp\left(-\frac{1}{2}x'Sx\right) dx \\ &+ \sum_{k=1}^p \lambda_k(\xi_{k\beta} - \xi_{k\alpha})^2 \int_{\mathbb{R}^p} (x_\alpha - x_\beta)^2 \exp\left(-\frac{1}{2}x'Sx\right) dx.\end{aligned}$$

Continuing,

$$
\begin{aligned}
&\int_{\mathbb{R}^p}(x'Tx-x'Sx)\exp\left(-\frac{1}{2}x'Sx\right)dx\\
=&2\sum_{k=1}^{p}\lambda_k(\xi_{k\beta}-\xi_{k\alpha})\Bigg[\xi_{k\alpha}\int_{\mathbb{R}^p}x_\alpha^2\exp\left(-\frac{1}{2}x'Sx\right)dx\\
&-\xi_{k\beta}\int_{\mathbb{R}^p}x_\beta^2\exp\left(-\frac{1}{2}x'Sx\right)dx\\
&+\sum_{\substack{\nu=1\\ \nu\neq\alpha}}^{p}\xi_{k\nu}\int_{\mathbb{R}^p}x_\nu x_\alpha\exp\left(-\frac{1}{2}x'Sx\right)dx\\
&-\sum_{\substack{\nu=1\\ \nu\neq\beta}}^{p}\xi_{k\nu}\int_{\mathbb{R}^p}x_\nu x_\beta\exp\left(-\frac{1}{2}x'Sx\right)dx\Bigg]\\
&+\sum_{k=1}^{p}\lambda_k(\xi_{k\beta}-\xi_{k\alpha})^2\Bigg[\int_{\mathbb{R}^p}x_\alpha^2\exp\left(-\frac{1}{2}x'Sx\right)dx\\
&+\int_{\mathbb{R}^p}x_\beta^2\exp\left(-\frac{1}{2}x'Sx\right)dx\\
&-2\int_{\mathbb{R}^p}x_\alpha x_\beta\exp\left(-\frac{1}{2}x'Sx\right)dx\Bigg].
\end{aligned}
$$

Let $S=\Phi'\Phi$ and $t=\Phi x$, then $\Phi=\Lambda^{\frac{1}{2}}\Gamma$ in virtue of $S=\Gamma'\Lambda\Gamma$. And $t_\alpha=\pm\sqrt{\lambda_\alpha}x_\alpha$ is from

$$
\begin{aligned}
x&=(\Lambda^{\frac{1}{2}}\Gamma)^{-1}t=\Gamma^{-1}\Lambda^{-\frac{1}{2}}t=\Gamma'\Lambda^{-\frac{1}{2}}t\\
&=(\xi_1',\cdots,\xi_p')\mathrm{diag}(\frac{1}{\sqrt{\lambda_1}},\cdots,\frac{1}{\sqrt{\lambda_p}})t\\
&=(\pm\frac{1}{\sqrt{\lambda_1}},\cdots,\pm\frac{1}{\sqrt{\lambda_p}})t.
\end{aligned}
$$

Thus, we obtain

$$
\begin{aligned}
&\int_{\mathbb{R}^p}x_\alpha^2\exp\left(-\frac{1}{2}x'Sx\right)dx\\
=&\frac{1}{|\det(\Phi)|}\int_{\mathbb{R}^p}\frac{t_\alpha^2}{\lambda_\alpha}\exp\left(-\frac{1}{2}t't\right)dt\\
=&\frac{1}{\sqrt{\det(S)}}\left(\prod_{\substack{\mu=1\\ \mu\neq\alpha}}^{p}\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}t_\mu^2\right)dt_\mu\right)\\
&\int_{-\infty}^{+\infty}\frac{t_\alpha^2}{\lambda_\alpha}\exp\left(-\frac{1}{2}t_\alpha^2\right)dt_\alpha\\
=&\frac{1}{\sqrt{\det(S)}}(\sqrt{2\pi})^{p-1}\frac{\sqrt{2\pi}}{\lambda_\alpha}=\frac{(\sqrt{2\pi})^p}{\sqrt{\det(S)}}\frac{1}{\lambda_\alpha}.
\end{aligned}
$$

Owing to the symmetry of the integral, we have

$$\int_{\mathbb{R}^p}x_\mu x_\nu e^{-\frac{1}{2}t't}dt=0\quad\text{for}\quad\mu\neq\nu.$$

With the aid of Lemma 4.2 and $\xi_{k\nu}=\pm\delta_{k\nu}$ (Kronecker's delta), we see that

$$
\begin{aligned}
&\mathbb{S}(\mathbb{P}_{x|S}\|\mathbb{P}_{x|T})\\
=&\frac{1}{\sqrt{(2\pi)^p\det(S)^{-1}}}\int_{\mathbb{R}^p}(x'Tx-x'Sx)\exp\left(-\frac{1}{2}x'Sx\right)dx\\
=&\frac{1}{\sqrt{(2\pi)^p\det(S)^{-1}}}\frac{(\sqrt{2\pi})^p}{\sqrt{\det(S)}}\sum_{k=1}^{p}\Bigg[2\lambda_k(\xi_{k\beta}-\xi_{k\alpha})\\
&\left(\frac{\xi_{k\alpha}}{\lambda_\alpha}-\frac{\xi_{k\beta}}{\lambda_\beta}\right)+\lambda_k(\xi_{k\beta}-\xi_{k\alpha})^2\left(\frac{1}{\lambda_\alpha}+\frac{1}{\lambda_\beta}\right)\Bigg]\\
=&2\lambda_\alpha(\mp1)\left(\frac{\pm1}{\lambda_\alpha}-\frac{0}{\lambda_\beta}\right)+\lambda_\alpha(\mp1)^2\left(\frac{1}{\lambda_\alpha}+\frac{1}{\lambda_\beta}\right)\\
&+2\lambda_\beta(\pm1)\left(\frac{0}{\lambda_\alpha}-\frac{\pm1}{\lambda_\beta}\right)+\lambda_\beta(\pm1)^2\left(\frac{1}{\lambda_\alpha}+\frac{1}{\lambda_\beta}\right)\\
=&\frac{\lambda_\alpha}{\lambda_\beta}+\frac{\lambda_\beta}{\lambda_\alpha}-2. \qquad (8)
\end{aligned}
$$

For the case II, again, we may assume that $1<\alpha<\beta<\gamma<\delta<p$ by re-ordering indices as needed be, then $\xi_k^*=\xi_k\Omega^{(\beta\gamma)}\Omega^{(\alpha\delta)}$ gives

$$
\begin{aligned}
\xi_k^*+\xi_k=&(2\xi_{k1},\cdots,\xi_{k\alpha}+\xi_{k\delta},\cdots,\xi_{k\beta}+\xi_{k\gamma},\cdots,\\
&\xi_{k\gamma}+\xi_{k\beta},\cdots,\xi_{k\delta}+\xi_{k\alpha},\cdots,2\xi_{kp}),\\
\xi_k^*-\xi_k=&(0,\cdots,\xi_{k\delta}-\xi_{k\alpha},\cdots,\xi_{k\gamma}-\xi_{k\beta},\cdots,\\
&\xi_{k\beta}-\xi_{k\gamma},\ldots,\xi_{k\alpha}-\xi_{k\delta},\ldots,0).
\end{aligned}
$$

The intermediate terms are the $\alpha^{th},\beta^{th},\gamma^{th}$ and $\delta^{th}$ positions, then we get

$$
\begin{aligned}
(\xi_k^*+\xi_k)x=&2\xi_kx+(\xi_{k\delta}-\xi_{k\alpha})(x_\alpha-x_\delta)\\
&+(\xi_{k\gamma}-\xi_{k\beta})(x_\beta-x_\gamma),
\end{aligned}
$$

and

$$
\begin{aligned}
(\xi_k^*-\xi_k)x=&(\xi_{k\delta}-\xi_{k\alpha})(x_\alpha-x_\delta)\\
&+(\xi_{k\gamma}-\xi_{k\beta})(x_\beta-x_\gamma).
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
&\mathbb{S}(\mathbb{P}_{x|S}\|\mathbb{P}_{x|T})\\
=&\frac{1}{\sqrt{(2\pi)^p\det(S)^{-1}}}\int_{\mathbb{R}^p}(x'Tx-x'Sx)\exp\left(-\frac{1}{2}x'Sx\right)dx\\
=&\frac{1}{\sqrt{(2\pi)^p\det(S)^{-1}}}\int_{\mathbb{R}^p}\left[\sum_{k=1}^{p}\lambda_k(\xi_k^*+\xi_k)x(\xi_k^*-\xi_k)x\right]\\
&\exp\left(-\frac{1}{2}x'Sx\right)dx\\
=&\frac{1}{\sqrt{(2\pi)^p\det(S)^{-1}}}\sum_{k=1}^{p}\Big[\lambda_k(I_1+I_2+I_3+I_4+I_5)\Big],
\end{aligned}
$$

where

$$\begin{aligned}
I_1 &= \int_{\mathbb{R}^p} 2\xi_k x(\xi_{k\delta} - \xi_{k\alpha})(x_\alpha - x_\delta) \exp\left(-\frac{1}{2}x'Sx\right) dx \\
&= \frac{2(\sqrt{2\pi})^p}{\sqrt{\det(S)}}(\xi_{k\delta} - \xi_{k\alpha})\left(\frac{\xi_{k\alpha}}{\lambda_\alpha} - \frac{\xi_{k\delta}}{\lambda_\delta}\right), \\
I_2 &= \int_{\mathbb{R}^p} 2\xi_k x(\xi_{k\gamma} - \xi_{k\beta})(x_\beta - x_\gamma) \exp\left(-\frac{1}{2}x'Sx\right) dx \\
&= \frac{2(\sqrt{2\pi})^p}{\sqrt{\det(S)}}(\xi_{k\gamma} - \xi_{k\beta})\left(\frac{\xi_{k\beta}}{\lambda_\beta} - \frac{\xi_{k\gamma}}{\lambda_\gamma}\right), \\
I_3 &= \int_{\mathbb{R}^p} (\xi_{k\delta} - \xi_{k\alpha})^2(x_\alpha - x_\delta)^2 \exp\left(-\frac{1}{2}x'Sx\right) dx \\
&= \frac{(\sqrt{2\pi})^p}{\sqrt{\det(S)}}(\xi_{k\delta} - \xi_{k\alpha})^2\left(\frac{1}{\lambda_\alpha} + \frac{1}{\lambda_\delta}\right), \\
I_4 &= \int_{\mathbb{R}^p} (\xi_{k\gamma} - \xi_{k\beta})^2(x_\beta - x_\gamma)^2 \exp\left(-\frac{1}{2}x'Sx\right) dx \\
&= \frac{(\sqrt{2\pi})^p}{\sqrt{\det(S)}}(\xi_{k\gamma} - \xi_{k\beta})^2\left(\frac{1}{\lambda_\beta} + \frac{1}{\lambda_\gamma}\right), \\
I_5 &= \int_{\mathbb{R}^p} (\xi_{k\delta} - \xi_{k\alpha})(\xi_{k\gamma} - \xi_{k\beta})(x_\alpha - x_\delta)(x_\beta - x_\gamma) \\
&\quad \exp\left(-\frac{1}{2}x'Sx\right) dx = 0.
\end{aligned}$$

Hence, we arrive at the following expression:

$$\begin{aligned}
&\mathbb{S}(\mathbb{P}_{x|S} \| \mathbb{P}_{x|T}) \\
=&\frac{1}{\sqrt{(2\pi)^p \det(S)^{-1}}} \frac{(\sqrt{2\pi})^p}{\sqrt{\det(S)}} \sum_{k=1}^{p} \Bigg\{ \lambda_k \Bigg[ 2(\xi_{k\delta} - \xi_{k\alpha}) \\
&\left(\frac{\xi_{k\alpha}}{\lambda_\alpha} - \frac{\xi_{k\delta}}{\lambda_\delta}\right) + 2(\xi_{k\gamma} - \xi_{k\beta})\left(\frac{\xi_{k\beta}}{\lambda_\beta} - \frac{\xi_{k\gamma}}{\lambda_\gamma}\right) + \\
&(\xi_{k\delta} - \xi_{k\alpha})^2\left(\frac{1}{\lambda_\alpha} + \frac{1}{\lambda_\delta}\right) + (\xi_{k\gamma} - \xi_{k\beta})^2\left(\frac{1}{\lambda_\beta} + \frac{1}{\lambda_\gamma}\right) \Bigg] \Bigg\} \\
=&\frac{\lambda_\alpha}{\lambda_\delta} + \frac{\lambda_\delta}{\lambda_\alpha} + \frac{\lambda_\beta}{\lambda_\gamma} + \frac{\lambda_\gamma}{\lambda_\beta} - 4. \qquad (9)
\end{aligned}$$

The last expression is due to that fact that $\xi_{k\nu} = \pm\delta_{k\nu}$ and applying the similar calculation in the last equality of obtaining Eq. (8).

We have thus proved the theorem. □