TENSOR NORMAL MODELS ON DIRECTED ACYCLIC GRAPHS

Abstract. Abstract here.

1. Introduction

Maximum Likelihood Estimation (MLE) is a fundamental problem in statistics, and recently it has been studied from the lens of algebra. The MLE problem is to find a point in a model that 'best' fits some data. The idea of 'fitting' a data is given by a *likelihood function*, and the problem of finding a point that 'best' fits the data translates to *maximizing* the likelihood function. A point that maximizes the likelihood function is called the *maximum likelihood estimate*.

Throughout we will use the calligraphic font $\mathcal{G}$ to denote a directed acyclic graph (DAG) and the normal font G to denote a group. $\mathcal{M}$, with some subscript containing an object, will be used to denote a model. For example, $\mathcal{M}_G$ is the model associated to a group G. PD_m will refer to the cone of symmetric $m \times m$ positive definite matrices.

2. Main Results

The main paper [1] deals mainly about Gaussian group models and section 5 talks about certain DAG models and their relationship with Gaussian group models. Anna Seigal suggested to deal with the model $\mathcal{M}(\mathcal{G}_1, \mathcal{G}_2) := \{\Psi_1 \otimes \Psi_2 : \Psi_i \in \mathcal{M}_{\mathcal{G}_i}\}$ when two DAGs $\mathcal{G}_1, \mathcal{G}_2$ are given, and raised the primary question that when are such models Gaussian group models, that is, when is $\mathcal{M}(\mathcal{G}_1, \mathcal{G}_2) = \mathcal{M}_G$ for some group G of matrices. For the one-parameter case, such models have been studied. Given a DAG $\mathcal{G}$, one considers the set of matrices

$$G(\mathcal{G}) = \left\{X \in GL_{|V(\mathcal{G})|} : X_{ij} = 0 \text{ for } i \neq j \text{ with } j \not\to i \text{ in } \mathcal{G}\right\}.$$

This is relevant in the 'good' case when $\mathcal{M}_{\mathcal{G}} = \mathcal{M}_{G(\mathcal{G})}$ as indicated in theorem 3.2.

The problem on $\mathcal{M}(\mathcal{G}_1, \mathcal{G}_2)$ naturally starts by considering

$$G(\mathcal{G}_1, \mathcal{G}_2) := \{\Psi_1 \otimes \Psi_2 : \Psi_i \in G(\mathcal{G}_i)\} = G(\mathcal{G}_1) \otimes G(\mathcal{G}_2).$$

It can be directly proven that

Proposition 2.1. *$G(\mathcal{G}_1, \mathcal{G}_2)$ is a group iff both $\mathcal{G}_i$ are TDAGs. If both $\mathcal{G}_i$'s are TDAGs, the model $M = \{\Psi_1 \otimes \Psi_2 : \Psi_i \in \mathcal{M}_{\mathcal{G}_i}\}$ is exactly $\mathcal{M}_{G(\mathcal{G}_1, \mathcal{G}_2)}$.*

My main contribution is to define a construction $\mathcal{G}_1 \otimes \mathcal{G}_2$ such that $G(\mathcal{G}_1) \otimes G(\mathcal{G}_2) \cong G(\mathcal{G}_1 \otimes \mathcal{G}_2)$. In fact such a construction is commutative (upto relabelling) and associative. Thus it extends to $\mathcal{G}_1 \otimes \cdots \otimes \mathcal{G}_n$ so that $\bigotimes_{i=1}^{n} G(\mathcal{G}_i) \cong G\left(\bigotimes_{i=1}^{n} \mathcal{G}_i\right)$. That is, this construction extends to the tensor normal models on n DAGs.

3. Background

3.1. **Maximum likelihood estimation.**

We will focus on multivariate Gaussian distributions of mean zero and covariance matrix Σ, whose density is given by

$$f_\Sigma(\boldsymbol{y}) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}\boldsymbol{y}^\intercal \Sigma^{-1} \boldsymbol{y}\right)$$

where $\boldsymbol{y} \in \mathbb{R}^m$ and $\Sigma \in \mathrm{PD}_m$. The corresponding $\Psi = \Sigma^{-1}$ will be its *concentration matrix*. A *Gaussian model* is a set of $m-$dimensional Gaussian distributions with mean zero. Such a model is given by a set of $m \times m$ symmetric positive definite covariance matrices. Equivalently they are also determined by a set of concentration matrices in PD_m. We will refer to this as the Gaussian model, instead of the set of densities themselves. So a Gaussian model $\mathcal{M}$ is just a subset of PD_m whose elements are to be thought of as concentration matrices.

A maximum likelihood estimate is a point Ψ in the model which maximizes the likelihood of observing some sample data point $\vec{\boldsymbol{Y}} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$. Mathematically, we want a $\Psi \in \mathcal{M}$ that maximizes the likelihood function $L_{\vec{\boldsymbol{Y}}}(\Psi) = \prod_{i=1}^{n} f_{\Psi^{-1}}(\boldsymbol{y}_i)$. Often, it is easier to maximize $l_{\vec{\boldsymbol{Y}}} := \log L_{\vec{\boldsymbol{Y}}}$

instead of $L_{\vec{\boldsymbol{Y}}}$ itself, and they have the same maximizers. Note that

$$\begin{aligned}
l_{\vec{\boldsymbol{Y}}}(\Psi) &= \log L_{\vec{\boldsymbol{Y}}}(\Psi) \\
&= \sum_{i=1}^{n} \left(-\frac{1}{2} \log\det(2\pi\Psi^{-1}) - \frac{1}{2} \boldsymbol{y}_i^{\intercal} \Psi \boldsymbol{y}_i \right) \\
\implies \frac{2}{n} l_{\vec{\boldsymbol{Y}}} &= -\log\det(2\pi\Psi^{-1}) - \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i^{\intercal} \Psi \boldsymbol{y}_i \\
&= -m \log(2\pi) + \log\det\Psi - \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i^{\intercal} \Psi \boldsymbol{y}_i \\
\implies \frac{2}{n} l_{\vec{\boldsymbol{Y}}} + m\log(2\pi) &= \log\det\Psi - \frac{1}{n} \sum_{i=1}^{n} \mathrm{Tr}\left(\boldsymbol{y}_i^{\intercal} \Psi \boldsymbol{y}_i \right) \\
&= \log\det\Psi - \mathrm{Tr}\left(\frac{1}{n} \sum_{i=1}^{n} \Psi \boldsymbol{y}_i^{\intercal} \boldsymbol{y}_i \right) \\
&= \log\det\Psi - \mathrm{Tr}\left(\Psi \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i^{\intercal} \right) \\
&= \log\det\Psi - \mathrm{Tr}\left(\Psi S_{\vec{\boldsymbol{Y}}} \right)
\end{aligned}$$

where $S_{\vec{\boldsymbol{Y}}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i^{\intercal}$ is the sample covariance matrix. This is clearly symmetric positive semi-definite. The above calculation shows that $\underset{\Psi \in \mathcal{M}}{\operatorname{argmax}} \left\{ l_{\vec{\boldsymbol{Y}}}(\Psi) \right\} = \underset{\Psi \in \mathcal{M}}{\operatorname{argmax}} \left\{ \log\det\Psi - \mathrm{Tr}\left(\Psi S_{\vec{\boldsymbol{Y}}} \right) \right\}$. Take $\ell_{\vec{\boldsymbol{Y}}}(\Psi) := \log\det\Psi - \mathrm{Tr}(\Psi S_{\vec{\boldsymbol{Y}}})$. Observe that $\ell_{\vec{\boldsymbol{Y}}}(\Psi)$ is invariant under similarity of Ψ and similarity of $S_{\vec{\boldsymbol{Y}}}$. Further $\Psi, S_{\vec{\boldsymbol{Y}}}$ are real symmetric matrices, hence have real eigenvalues and are diagonalizable. This means that if $\{\lambda_i\}_{i=1}^{m} \subseteq \mathbb{R}^{>0}$ and $\{\mu_i\}_{i=1}^{m} \subseteq \mathbb{R}^{\geq 0}$ are eigenvalues of Ψ and $S_{\vec{\boldsymbol{Y}}}$ respectively, then $\ell_{\vec{\boldsymbol{Y}}}(\Psi) = \sum \log\lambda_i - \sum \lambda_i \mu_i$. Note that if $S_{\vec{\boldsymbol{Y}}}$ is invertible, then

$$\begin{aligned}
\ell_{\vec{\boldsymbol{Y}}}(\Psi) &= \log\det\Psi - \operatorname{Tr}\left(\Psi S_{\vec{\boldsymbol{Y}}}\right) \\
&= \log\det\Psi - \log\det S_{\vec{\boldsymbol{Y}}}^{-1} + \log\det S_{\vec{\boldsymbol{Y}}}^{-1} \\
&\qquad - \operatorname{Tr}\left(\Psi S_{\vec{\boldsymbol{Y}}}\right) + \operatorname{Tr}\mathbf{1}_m - \operatorname{Tr}\left(S_{\vec{\boldsymbol{Y}}}^{-1} S_{\vec{\boldsymbol{Y}}}\right) \\
&= \sum \log(\lambda_i\mu_i) + \log\det S_{\vec{\boldsymbol{Y}}}^{-1} - \sum \lambda_i\mu_i + m - \operatorname{Tr}\left(S_{\vec{\boldsymbol{Y}}}^{-1} S_{\vec{\boldsymbol{Y}}}\right) \\
&= \sum \left[\log(\lambda_i\mu_i) - \lambda_i\mu_i\right] + m + \ell_{\vec{\boldsymbol{Y}}}\left(S_{\vec{\boldsymbol{Y}}}^{-1}\right) \\
&\le \sum_{i=1}^{m}(-1) + m + \ell_{\vec{\boldsymbol{Y}}}\left(S_{\vec{\boldsymbol{Y}}}^{-1}\right) = \ell_{\vec{\boldsymbol{Y}}}\left(S_{\vec{\boldsymbol{Y}}}^{-1}\right)
\end{aligned}$$

If $S_{\vec{\boldsymbol{Y}}}$ is singular, assume WLOG $\mu_1 = 0$, then $\ell_{\vec{\boldsymbol{Y}}}(\Psi) = \log\lambda_1 + \sum_{i\ge 2}(\log\lambda_i - \lambda_i\mu_i)$ diverges to $+\infty$ as $\lambda_1 \to \infty$.

3.2. **Gaussian group models.** The *Gaussian group model* given by a group $G \subseteq \mathrm{GL}(\mathbb{R}^n)$ is the multivariate Gaussian model comprising all densities of mean zero and concentration matrices given by

$$\mathcal{M}_G := \left\{X^\intercal X : X \in G\right\}.$$

The importance of such models is that finding an MLE is equivalent to an optimization problem. This is seen by the following calculation

$$-\ell_{\vec{\boldsymbol{Y}}}\left(X^\intercal X\right) = \frac{1}{n}\sum \operatorname{Tr}\left(\left(X\boldsymbol{y}_i\right)^\intercal X\boldsymbol{y}_i\right) - \log\det\left(X^\intercal X\right) \tag{1}$$

$$= \frac{1}{n}\left|\left|X\cdot\vec{\boldsymbol{Y}}\right|\right|_2^2 - \log(\det X)^2 \tag{2}$$

So maximizing $\ell_{\vec{\boldsymbol{Y}}}\left(X^\intercal X\right)$ is equivalent to minimizing $\frac{1}{n}\left|\left|X\vec{\boldsymbol{Y}}\right|\right|_2^2 - \log\det\left(X^\intercal X\right)$ which consists of minimizing norms. The exact statement is captured in the following theorem in [1, Proposition 3.4]:

Proposition 3.1. *Let $\vec{\boldsymbol{Y}} \in V^n$ be a tuple of samples. If the group $G \subseteq \mathrm{GL}(V)$ is closed under non-zero scalar multiples, the supremum of the log-likelihood $\ell_{\vec{\boldsymbol{Y}}}(\Psi)$ over $\mathcal{M}_G$ is the double infimum*

$$-\inf_{\lambda>0}\left(\frac{\lambda}{n}\left(\inf_{H\in G_{\mathrm{SL}}^{\pm}}\left|\left|H\cdot\vec{\boldsymbol{Y}}\right|\right|^2\right) - \dim(V)\log\lambda\right).$$

The MLEs, if they exist, are the matrices $\dfrac{n \dim V}{\left|\left|H \cdot \vec{\boldsymbol{Y}}\right|\right|^2} H^\intercal H$*, where* H *minimizes* $\left|\left|H \cdot \vec{\boldsymbol{Y}}\right|\right|$ *under the action of* $G_{\mathrm{SL}}^{\pm}$ *on* V^n*.*

In the above, $G_{\mathrm{SL}}^{\pm} = \{X \in G : \det X = \pm 1\}$.

What the above proposition says is that $\ell_{\vec{\boldsymbol{Y}}}\left(X^\intercal X\right)$ can be maximized in two steps:

(1) Minimize the norm $\left|\left|H \cdot \vec{\boldsymbol{Y}}\right|\right|^2$ over $H \in G_{\mathrm{SL}}^{\pm}$.

(2) Find a scalar $\mu \in \mathbb{R}$ so that $X \coloneqq \mu H$ mimimizes $-\ell_{\vec{\boldsymbol{Y}}}\left(X^\intercal X\right)$ in eq. (2).

Proof. We want to minimize the function $f : G \to \mathbb{R}$ given by

$$f(X) = \frac{1}{n}\left|\left|X \cdot \vec{\boldsymbol{Y}}\right|\right|_2^2 - \log\left(\det X\right)^2.$$

Let $m = \dim V$. Observe that $f|_{G_{\mathrm{SL}}^{\pm}}$ determines f completely. This is because for any $X \in G$, take $Z \coloneqq \dfrac{1}{\mu_X} X$, where $\mu_X = (|\det X|)^{\frac{1}{m}} \in \mathbb{R}_{>0}$. Clearly $\det Z = \frac{\det X}{|\det X|} = \pm 1$. Then

$$\begin{aligned} f(X) &= \frac{\mu_X^2}{n}\left|\left|Z \cdot \vec{\boldsymbol{Y}}\right|\right|_2^2 - \log\left(\left(\mu_X^m \det Z\right)^2\right) \\ &= \frac{\mu_X^2}{n}\left|\left|Z \cdot \vec{\boldsymbol{Y}}\right|\right|_2^2 - 2m \log \mu_X \\ &= \mu_X^2 f(Z) - 2m \log \mu_X \end{aligned}$$

Note that for $K > 0$, the function $s \mapsto sK - \log s$ minimizes at $s = \frac{1}{K}$ giving a min value of $1 + \log K$, the latter being an increasing function of K. Thus minimizing $f(X)$ is equivalent to first minimizing the norm in the orbit of $G_{\mathrm{SL}}^{\pm}$ and then minimizing the overall expression with the previous minima. In other words,

$$\inf_{X \in G} f(X) = \inf_{\mu > 0}\left(\mu^2 \cdot \inf_{Z \in G_{SL}^{\pm}}\left|\left|Z \cdot \vec{\boldsymbol{Y}}\right|\right|_2^2 - m \log \mu^2\right).$$

Replacing $\lambda = \mu^2$ gives the expression in the expression in the proposition. The MLE, if it exists, is given by the point $\hat{X} = \hat{\mu}\hat{Z}$, where

$\hat{Z}$ minimizes $\left|\left|Z \cdot \vec{\boldsymbol{Y}}\right|\right|_2^2$ and $\hat{\mu} = \dfrac{\sqrt{mn}}{\left|\left|\hat{Z} \cdot \vec{\boldsymbol{Y}}\right|\right|_2}$, which corresponds to the input $\hat{\Psi} = \hat{X}^\intercal \hat{X} = \hat{\mu}^2 \hat{Z}^\intercal \hat{Z}$. ■

3.3. **Transitive DAGs.** The relevance of transitive DAGs in the statistical context was introduced in [1, §5]. We briefly introduce the relevant details here. Every DAG $\mathcal{G}$ has a model associated to them given by

$$\mathcal{M}_\mathcal{G} := \left\{(\mathbf{1} - \Lambda)^\intercal \Omega^{-1}(\mathbf{1} - \Lambda) : \Lambda, \Omega \in \mathbb{R}^{m_1 \times m_i}, \Lambda_{ij} \neq 0 \implies j \to i, \Omega \text{ diagonal and PD}\right\}.$$

Note that taking $\Omega = \mathbf{1}, \Lambda = \mathbf{0}$ forces $\mathbf{1} \in \mathcal{M}_\mathcal{G}$. These matrices $\Psi \in \mathcal{M}_\mathcal{G}$ are to be thought of as concentration matrices, so they define a model $\{f_{\Psi^{-1}}\}$ (which is just a collection of probability densities) given by

$$f_\Sigma(y) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2} y^\intercal \Sigma^{-1} y\right)$$

where $\Sigma = \Psi^{-1}$.

$\mathcal{G}$ is said to be a *transitive* DAG (TDAG) if $i \to k$ is an edge in $\mathcal{G}$ whenever $i \to j, j \to k$ are. They turn out to be the 'good' DAGs that help relate these models to Gaussian group models due to the following proposition in [1]:

Theorem 3.2. *$G(\mathcal{G})$ is a group iff $\mathcal{G}$ is a TDAG. In such a case, $\mathcal{M}_\mathcal{G} = \mathcal{M}_{G(\mathcal{G})}$.*

3.4. **ss.**

References

[1] Carlos Améndola, Kathlén Kohn, Philipp Reichenbach, and Anna Seigal. Invariant theory and scaling algorithms for maximum likelihood estimation. *SIAM Journal on Applied Algebra and Geometry*, 5(2):304–337, Jan 2021.