科研費シンポジウム「統計的推測理論とその応用—正則から非正則へ—」
2000 年 9 月 4 日〜6 日

錐型の特異点を持つモデルにおける最尤推定量の漸近的挙動

Asymptotic properties of the maximum likelihood estimator in a model with a conic singularity

福水 健次
統計数理研究所
〒 106-8569　港区南麻布 4-6-7
Kenji Fukumizu
E-mail: fukumizu@ism.ac.jp

Abstract

本論分は、真のパラメータが識別不能な場合の最尤推定量の漸近的挙動について論じる　特に、尤度比や KL-divergence がサンプルサイズ n に対してどのようなオーダーを持つかを考察する。パラメータの識別不能性は、有限混合モデル、縮小ランクなどの多くの重要な統計モデルに存在しているが、本論文では、特にニューラルネットを中心に論じる。Hartigan のアイデアに従い、識別不能性を、確率密度関数全体の空間の中での、モデルの錐型特異点として定式化し、最尤推定における尤度比を、関数族の上のガウス過程の極大値を用いて表現する。これを用いて、その関数族がある条件を満たせば、尤度比と KL-divergence は漸近的に一致し、$O(1/n)$ のオーダーを持つことを示す。識別不能なモデルの中には通常の $O(1/n)$ よりも大きいオーダーを持つものがあることが知られているが、そのような大きいオーダーを持つための関数族の十分条件を示す。これらの結果を用いてニューラルネットワークの尤度比について論じる。

1 Introduction

This paper discusses the asymptotic behavior of the maximum likelihood estimator (MLE) under the condition that the true parameter is unidentifiable. The asymptotic of MLE is an important problem in statistical estimation theory, and the asymptotic normality under some regularization conditions are well known ([1]). However, if the true parameter is of dimension larger than one, the Fisher information matrix at the true parameter is singular, and the asymptotic normality is no longer satisfied. The asymptotic behavior of MLE in such unidentifiable situations has not been clarified completely.

We formulate the problem of unidentifiability as a conic singularity ([2]) in the set of a statistical model, embedded in the space of all the probability density functions. In this formulation, the likelihood ratio of the MLE, with the true probability at the singularity, can be described by the maximum of a Gaussian process over the unit vectors in the tangent cone. This Gaussian process shows very different behavior depending on the functional property of the tangent cone. One of the interesting feature is the order of the likelihood ratio as the number of samples n goes to infinity. A model satisfying the regularity condition of usual asymptotic theory has the likelihood ratio of the order $O(1/n)$. However, a larger order has been reported in some models. Hartigan ([3]) discusses the normal mixture models. In neural networks, the order $O(\log n/n)$ has been derived in unidentifiable cases ([4]). A useful sufficient condition of larger order than $O(1/n)$ will be given in the term of tangent cone. I will further derive the order of the likelihood ration for some neural networks models, with the true probability at the singularity, analyzing the functional properties of the tangent cone.

2 Unidentifiability and Locally Conic Models

2.1 Preliminaries

Let $(\mathcal{Z}, \mathcal{B}, \mu)$ be a measure space. A *statistical model* $S = \{f(z;\theta) \mid \theta \in \Theta\}$ is a family of probability density functions on $(\mathcal{Z}, \mathcal{B}, \mu)$, where the parameter space Θ is a domain in the d-dimensional Euclidean space $\mathbb{R}^d$. We assume that $f(z;\theta) > 0$ for all z and θ, and differentiable on θ for each $z \in \mathcal{Z}$. Suppose the probability distribution of i.i.d. random variables $Z_1, Z_2, \ldots, Z_n$ is $f(z)\mu$ with the probability density function $f(z) > 0$. Given the random variables, the *likelihood ratio* of the model S with respect to $\{Z_i\}_{i=1}^n$ is defined by

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(Z_i;\theta)}{f(Z_i)}. \tag{1}$$

] Note that $L_n(\theta)$ is normalized by $1/n$, so that it can be compared with the Kullback-Leibler divergence, which will be defined later. We consider the *maximum likelihood estimator* (MLE) $\hat{\theta}$ that attains the maximum of the likelihood ratio, if it exists. The following equations hold;

$$L_n(\hat{\theta}) = \sup_{\theta\in\Theta} L_n(\theta) = \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(Z_i;\theta)}{f(Z_i)}. \tag{2}$$

For a density $f(z;\theta)$ in the model S, we define the *Kullback-Leibler divergence* of $f(z;\theta)$ from $f(z)$ by

$$D(\theta) = \int f(z) \log \frac{f(z)}{f(z;\theta)} d\mu(z). \tag{3}$$

The main topics of this paper is the behavior of the likelihood ratio and the Kullback-Leibler divergence of the maximum likelihood estimator under the asymptotic assumption.

2.2 Unidentifiability of the true parameter

Throughout this paper, the true probability distribution $f(z)\mu$ is assumed to be included in the model $\{f(z;\theta) \mid \theta \in \Theta\}$. Therefore, there exists $\theta_0 \in \Theta$ such that $f(z;\theta_0) = f(z)$. In some cases, the true parameter θ_0 is not unique. The usual view of asymptotic convergence to a single true parameter does not hold in such cases. In fact, the asymptotic theory assumes the uniqueness of the true parameter in its regularity conditions.

If the set of true parameter to give $f(z)$ is of dimension more than 0, we say the true parameter is unidentifiable. There are many statistical models with unidentifiability. A famous one is seen in finite mixture models. Let $g(z;a)$ be a probability density function on $\mathcal{Z}$ with variable parameter a, and $f(z;a_1,a_2,b)$ be a mixture model defined by

$$f(z;a_1,a_2,b) = b\, g(z;a_1) + (1-b)\, g(z;a_2), \tag{4}$$

where $b \in [0,1]$. Suppose the true density is given by $g(z;a_0)$ for some a_0, then, the set of parameters to give $g(z;a_0)$ is $\{(a_1,a_2,b) \mid a_1 = a_2 = a_0,\ b : \text{free}\} \cup \{(a_1,a_2,b) \mid b = 0, a_2 = a_0,\ a_1 : \text{free}\} \cup \{(a_1,a_2,b) \mid b = 1, a_1 = a_0,\ a_2 : \text{free}\}$, which is high dimensional. Hartigan ([3]) discusses the Gaussian mixture with two components. Besides mixture models, the reduced rank problems ([5]) and the change point problem ([6]) are examples of models with unidentifiability. Feed-forward neural network models, such as multilayer perceptrons ([7]), also have unidetifiability. We will mainly discuss multilayer perceptrons as an example.

The main concern of this paper is to investigate how the likelihood ratio or the Kullback-Leibler divergence asymptotically behaves if the true parameters are unidentifiable. As a comparison, if the true function is identifiable, under some regularity conditions, the asymptotic distribution of the likelihood ratio and the Kullback-Leibler divergence is well known. They have the same value in the leading term;

$$L_n(\hat{\theta}) = D(\hat{\theta}) + o_p(1/n), \tag{5}$$

and their limiting distribution is given by

$$nL_n(\hat{\theta}) \xrightarrow[n\to\infty]{} \chi^2_d \quad \text{in law}, \tag{6}$$

where χ^2_d denotes the chi-square distribution of freedom d.

2.3 Conic singularity

Following Dacunha-Castelle & Gassiat ([2]), with some modification, we utilize a conic singularity to formulate the unidentifiability.

Let $\Theta \subset \mathbb{R}^d$ be an open set, and S be a statistical model $\{f(z;\theta) \mid \theta \in \Theta\}$. Suppose $f_0(z)$ is an element in S. A parameter $\theta \in \Theta$ is decomposed as $\theta = (\alpha, \beta)$ for $\alpha \in \mathbb{R}^{d-1}$ and $\beta \in \mathbb{R}$. The statistical model S is called locally conic at f_0 if the following conditions are satisfied;

1. $f(z;\theta)$ is C^∞ function of θ for almost every z.
2. Let $\Theta_0 = \Theta \cap (\mathbb{R}^{d-1} \times \{0\})$, $A_0 = \{\alpha \in \mathbb{R}^{d-1} \mid (\alpha, 0) \in \Theta_0\}$, and $\Theta(\alpha) = \Theta \cap (\{\alpha\} \times \mathbb{R})$ for each $\alpha \in A_0$. Then,
$$\Theta = \bigcup_{\alpha \in A_0} \Theta(\alpha). \tag{7}$$
3. The set of the parameters to give f_0 is Θ_0; that is,
$$f(z;(\alpha,\beta))\mu = f_0(z)\mu \iff \beta = 0. \tag{8}$$
4. $\frac{\partial}{\partial\beta} \log f(z;\alpha,\beta)$ is in $L^2(f_{(\alpha,\beta)}\mu)$ and
$$\left\| \frac{\partial}{\partial\beta} \log f(z;\alpha,0) \right\|_{L^2(f_0\mu)} = 1 \tag{9}$$
for all $\alpha \in A_0$.

If the set Θ_0 is not a single point, the parameter giving f_0 is not identifiable. Geometrically, a locally conic model S is a d-dimensional set with a singularity at f_0 in the space of probability density functions. The score function of the submodel $S_\alpha = \{f(z;\theta) \mid \theta \in \Theta(\alpha)\}$ at the origin,

$$v_\alpha(z) = \frac{\partial f(z;(\alpha,0))}{\partial\beta}, \tag{10}$$

can be looked as a unit tangent vector in the direction of S_α. The family of score functions $C = \{v_\alpha\}$ generates the tangent cone at the singularity f_0. We call C as the *basis of the tangent cone*. This view of tangent vectors can be rigorously formulated if S is included in a maximal exponential model ([8]), which is an infinite dimensional Banach manifold. The basis of the tangent cone C has a key importance in the following discussion. In the definition, we require only that the functions in C are in $L^2(f_0\mu)$. They are not necessarily real tangent vectors in the Banach manifold.

2.4 Neural networks

A feed-forward neural network model is an example of a model with unidentifiability. We mainly discuss multilayer perceptrons ([7]) in later sections. The *multilayer perceptron* model with H hidden units is defined by a family of functions

$$\varphi(x;\theta) = \sum_{j=1}^{H} b_j\, s(a_j x + c_j) + d, \tag{11}$$

where $x \in \mathcal{X} = \mathbb{R}$, $s(t) = \tanh(t)$, and $\theta = (a_1, c_1, b_1, \dots, a_H, c_H, b_H, d)^T$. We discuss only one-dimensional input and output for simplicity.

We can regard learning in neural networks as statistical estimation. Assume a probability $Q = q(x)dx$ on $\mathcal{X}$ for the distribution of the input sample X_i, and a conditional probability density function $r(y \mid u)$ of $y \in \mathcal{Y} = \mathbb{R}$ given $u \in \mathbb{R}$. Throughout this paper we assume the existence of the second moment for Q and $r(y|u)dy$. Define a statistical model by

$$p(z;\theta) = r(y \mid \varphi(x;\theta))q(x), \tag{12}$$

where $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

Useful choices of $r(y \mid u)$ are the additive Gaussian noise model

$$r(y \mid u) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{-\frac{1}{2\sigma^2}(y-u)^2\Big\} \tag{13}$$

for continuous y, and the binomial distribution model

$$r(y \mid u) = u^y(1-u)^{1-y} \tag{14}$$

for binary output y, which often appears in classification problems. The maximum likelihood estimation gives the least mean squares

$$\min_{\theta\in\Theta} \sum_{i=1}^{n} (Y_i - \varphi(X_i;\theta))^2 \tag{15}$$

for the former example, and the cross-entropy loss function

$$\min_{\theta\in\Theta}\sum_{i=1}^{n}\{-Y_i\log\varphi(X_i;\theta)-(1-Y_i)\log(1-\varphi(X_i;\theta))\} \tag{16}$$

for the latter example.

The true parameter can be unidentifiable in the multilayer perceptron model. We see it using the simplest case. Suppose we have the multilayer perceptron model with 2 hidden units, and the true function $\varphi_0(x)$ can be given by a perceptron with only one hidden unit. If $\varphi_0(x) = b_0\tanh(a_0x)$, then for any parameter θ in the set $\{(a_1,c_1,b_1,a_2,c_2,b_2,d)\in\Theta \mid a_1 = a_0, b_1 = b_0, c_1 = 0, b_2 = 0, d = 0\}$ and $\{(a_1,c_1,b_1,a_2,c_2,b_2,d)\in\Theta \mid a_1 = a_0, b_1 = b_0, c_1 = 0, a_2 = 0, b_2\tanh(c_2)+d = 0\}$ the function $\varphi(x;\theta)$ equals to the true function [1] We can see that the set of true parameters is a high dimensional subset in the parameter space. It is known if a function can be realized by a network with smaller number of hidden units than the model, the set of parameters to give the function is high dimensional set ([9],[10],[11]).

2.5 Multilayer perceptron as a locally conic model

Many statistical models with unidentifiable parameters can be described by locally conic models. Dacunha-Castelle and Gassiat ([2]) discusses a finite mixture model as a locally conic model, while the one-dimensional submodel is defined on the half line $[0,\infty)$ unlike ours. We will show that the unidentifiability of neural networks is formulated as a conic singularity.

Suppose we have the multilayer perceptrons with H hidden units. Let $K\in\mathbb{N}$ be less than H, and $\varphi_0(x)$ be a function realizable by a multilayer perceptron with K hidden units. In the parameter space of the model with H hidden units, the parameter to give the function $\varphi_0(x)$ is unidentifiable, because it is high dimensional ([9],[12],[11]). We can rewrite this unidentifiability by a conic singularity using a new parameterization.

For simplicity, we consider only multilayer perceptrons without bias terms. Then, the model is defined by a family functions:

$$\varphi(x;\theta) = \sum_{j=1}^{H} b_j\, s(a_jx), \tag{17}$$

[1] These two subsets do not give all the parameters to realize $\varphi_0(x)$. The whole set of the true parameters is shown in [11].

where $\theta = (a_1, \ldots, a_H, b_1, \ldots, b_H)^T \in \mathbb{R}^{2H}$. The existence of the bias terms influences much on the functional properties of the model. However, we choose this simpler form to avoid the technical difficulties.

Let $\Theta_H^* = \{\theta = (a_1, \ldots, a_H, b_1, \ldots, b_H) \in \mathbb{R}^{2H} \mid a_j \neq 0,\ b_j \neq 0\ (1 \leq j \leq H),\ |a_j| \neq |a_h|\ (1 \leq j < h \leq H)\}$ be the parameter space of the multilayer perceptrons with H hidden units. Note that we eliminate the parameters which correspond functions realizable by a smaller-sized network. This modification does not matter in discussing the maximum likelihood estimation, because the maximum likelihood estimator lies in Θ^* with probability one. For a parameter in Θ_H^*, it is known ([12]) that the functions $s(a_j x)$ and $s'(a_j x)x$ $(1 \leq j \leq H)$ are linearly independent.

Given a function $\varphi(x) = \sum_{k=1}^{K} b_k^0\, s(a_k^0 x)$, $((a_k^0, b_k^0) \in \Theta_K^*)$, we slightly modify the parameter space as $\Theta_H^{**} = \{\theta \in \Theta_H^* \mid |a_j| \neq |a_k^0|\ (1 \leq k \leq K, K+1 \leq j \leq H)\}$, and introduce a new parameterization by

$$
\begin{aligned}
&\beta = \mathrm{sgn}(b_{K+1})\sqrt{b_{K+1}^2 + \cdots + b_H^2}, \\
&\xi_k = \frac{a_k - a_0}{\beta}, \quad (1 \leq k \leq K), \qquad \xi_j = a_j, \quad (K+1 \leq j \leq H), \\
&\eta_k = \frac{b_k - b_0}{\beta}, \quad (1 \leq k \leq K), \qquad \eta_j = \frac{b_j}{\beta}, \quad (K+1 \leq j \leq H). \qquad (18)
\end{aligned}
$$

for $\theta \in \Theta_H^{**}$. Define a new parameter space Π_H by

$$
\begin{aligned}
\Pi_H = \{\omega = (\xi_1, \ldots, \xi_H, \eta_1, \ldots, \eta_H, \beta) \mid\ & a_k^0 + \beta\xi_k \neq 0\ (1 \leq k \leq K), \\
& \xi_j \neq 0\ (K+1 \leq j \leq H),\ |a_k^0 + \beta\xi_k| \neq |a_h^0 + \beta\xi_h|\ (1 \leq k < h \leq H), \\
& |a_k^0 + \beta\xi_k| \neq |\xi_j|\ (1 \leq k \leq K, K+1 \leq j \leq H), \\
& |\xi_j| \neq |\xi_i|\ (K+1 \leq j < i \leq H),\ |\xi_j| \neq |a_k^0|\ (1 \leq k \leq K, K+1 \leq j \leq H), \\
& b_k^0 + \beta\eta_k \neq 0\ (1 \leq k \leq K),\ \sum_{j=K+1}^{H} \eta_j^2 = 1,\ \eta_j \neq 0\ (K+1 \leq j \leq H), \\
& \eta_{K+1} > 0,\ \beta \in \mathbb{R}\} \qquad (19)
\end{aligned}
$$

and $\Pi_H^{**} = \{\omega \in \Pi_H \mid \beta \neq 0\}$. Rewrite the multilayer perceptron using this parameterization;

$$
\psi(x;\omega) = \sum_{k=1}^{K} (b_k^0 + \beta\eta_k) s((a_k^0 + \beta\xi_k)x) + \sum_{j=K+1}^{H} \beta\eta_j s(\xi_j x). \qquad (20)
$$

It is easy to see that the Π_H^{**} and Θ_H^{**} are diffeomorphic by the above correspondence, and $\varphi(x;\theta) = \psi(x;\omega)$ for the corresponding $\theta \in \Theta_H^{**}$ and $\omega \in \Pi_H^{**}$.

We write $\omega = (\alpha, \beta)$, summarizing $(\xi_1, \ldots, \eta_H)$ by α. By the fact $(a_1^0, \ldots, a_K^0, b_1^0, \ldots, b_K^0) \in \Theta_K^*$ and $|\xi_j| \neq |a_k^0|$, we can show that $\Pi_H^{**} = \cup_{(\alpha,0)\in\Pi_{H,0}} \Pi_H(\alpha)$. Consider the family of functions $\{\psi(x;\omega) \mid \omega \in \Pi_H\}$. We can see that $\psi(x;\omega) = \varphi(x)$ if and only if $\omega \in \Pi_{H,0}$; that is, $\beta = 0$. The sufficiency is trivial. For the necessity, because both the sets $\{s(\xi_j x), s(a_k^0 x)\}$ and $\{s(\xi_j x), s((a_k^0 + \beta\zeta_k)x)\}$ are linearly independent, we see that the coefficients of $s(\xi_j x)$ must be zero to realize $\psi(x;\omega) = \varphi_0(x)$. This implies $\beta = 0$.

The basis of the tangent cone is essentially determined by the following partial derivatives;

$$\frac{\partial\psi(x;(\alpha,0))}{\partial\beta} = \sum_{j=K+1}^{H} \eta_j s(\xi_j x) + \sum_{k=1}^{K} \eta_k s(a_k^0 x) + \sum_{k=1}^{K} b_k^0 \xi_k s'(a_k^0 x) x. \quad (21)$$

Let $q(x)$ be a p.d.f. of x, such that $q(x)$ is absolute continuous with respect to the Lebesgue measure on $\mathbb{R}$. Let $r(y|u)$ be a conditional p.d.f. of y given u, such that $r(y|u_1)dy \neq r(y|u_2)dy$ for different u_1 and u_2, and the Fisher information $I(u)$ is positive and finite; $0 < I(u) = \int \big(\frac{\partial \log r(y|u)}{\partial u}\big)^2 r(y|u)dy < \infty$. We assume that $I(u)$ is bounded on a bounded interval in $\mathbb{R}$. Let $S_H = \{f(x,y;\omega) \mid \omega \in \Pi_H\}$ be a statistical model defined by $f(x,y;\omega) = r(y|\psi(x;\omega))q(x)$. The model S_H consists of probability density functions corresponding to $\varphi_0(x)$ and the functions realized by multilayer perceptrons with H hidden units and not by a smaller-sized network. The function $f_0(x,y)$ be a density defined by $\varphi_0(x)$, that is, $f_0(x,y) = r(y|\varphi(x))q(x)$. We have the following proposition;

Proposition 1. *The statistical model of multilayer perceptrons with H hidden units S_H is locally conic at a point f_0, which corresponds to a function realized by a network with K hidden units ($0 \leq K < H$).*

Proof. From what we have seen, the model S satisfies the conditions 1, 2, and 3 in the definition of a locally conic model. For the condition 4, let $N(\alpha)$ be the $L^2(f_0(x,y)dxdy)$-norm of $\frac{\partial}{\partial\beta}\log f(x,y;(\alpha,0))$. We have

$$\begin{aligned} N(\alpha)^2 &= \int\int r(y|\varphi_0(x))q(x)\Big(\frac{\partial r(y|\varphi_0(x))}{\partial u}\frac{\partial\psi(x;(\alpha,0))}{\partial\beta}\Big)^2 dxdy \\ &= \int I(\varphi(x))\Big(\frac{\partial\psi(x;(\alpha,0))}{\partial\beta}\Big)^2 q(x)dx. \end{aligned} \quad (22)$$

Because $\varphi_0(x)$ is a bounded function, $I(\varphi_0(x))$ is bounded and non-zero. The function $\big(\frac{\partial}{\partial u}r(y|\varphi(x))\frac{\partial}{\partial\beta}\psi(x;(\alpha,0))\big)^2$ is also bounded from eq.(21).

Thus, the integral eq.(22) is finite. Because $s(\xi_j x)$, $s(a_k^0 x)$, and $s'(a_k^0 x)x$ are linearly independent (see [12]), the partial derivative $\frac{\partial}{\partial\beta}\psi(x;(\alpha,0))$ is not constant zero. Therefore, $0 < N(\alpha) < \infty$ for all $\alpha \in A_0$. Using $N(\alpha)\beta$ instead of β, we have the normalized tangent vectors at $f_0(x,y)$. □

We discuss a special case in which the noise model $r(y|u)$ is an exponential family and the true function $\varphi_0(x)$ is constant zero. Let $\tilde{v}_\alpha$ be the tangent vector $\frac{\partial}{\partial\beta}f(x,y;(\alpha,0))$, without the normalization of the parameter β. As we see in the above proof, it is given by $\tilde{v}_\alpha = \frac{\partial}{\partial u}r(y|\varphi(x))\frac{\partial}{\partial\beta}\psi(x;(\alpha,0))$. Suppose that the conditional probability density $r(y|u)$ is given by an exponential family $r(y|u) = \exp\{y\kappa(u)+\tau(y)-\zeta(u)\}$, where $\kappa(u)$ is an invertible smooth function, and assume that $\int yr(y|u)dy = u$. This assumption is natural for the noise model. In this case, the score function is given by

$$\frac{\partial \log r(y|u)}{\partial u} = \frac{\partial\kappa(u)}{\partial u}(y-u), \tag{23}$$

and the tangent vectors are

$$\tilde{v}_\alpha = \frac{\partial\kappa(\varphi_0(x))}{\partial u}(y-\varphi_0(x))\frac{\partial\psi(x;(\alpha,0))}{\partial\beta}. \tag{24}$$

Moreover, if $\varphi_0(x) = 0$ (constant zero function), the tangent vectors are given by

$$\tilde{v}_\alpha = \kappa'(0)\ y\ \Big(\sum_{j=1}^{H}\eta_j s(\xi_j x)\Big), \tag{25}$$

which form the function class of multilayer perceptrons with H hidden units multiplied by y.

3 Maximum likelihood estimation in locally conic models

3.1 Maximum likelihood estimation as a supremum of a random process

Let $S = \{f(z;(\alpha,\beta)) \mid (\alpha,\beta) \in \Theta\}$ be a statistical model, which is locally conic at $f_0 \in S$. Suppose $Z_1, Z_2, \ldots, Z_n$ are i.i.d. random variables with the law $f_0\mu$. For each α satisfying $\alpha \in A_0$, define a submodel $S_\alpha = \{f(z;(\alpha,\beta)) \mid \beta \in \Theta(\alpha)\}$ is a smooth, one-dimensional statistical model with a variable

parameter β, and the Fisher information at the origin equal to one. Consider the maximum likelihood estimator $\hat{\beta}_\alpha$ in S_α, then, the maximum likelihood estimator in S is given by

$$\sup_{\theta\in\Theta} L_n(\theta) = \sup_{\alpha} L_n(\alpha, \hat{\beta}_\alpha). \tag{26}$$

Fix α and concentrate the maximum likelihood estimation in S_α for a while. The true parameter in $\Theta(\alpha)$ is 0. Assume that each submodel satisfy the regularity conditions of the asymptotic efficiency. A set of conditions is found in Sen and Singer ([13], Theorem 5.2.1), which shows weaker conditions than the famous ones by Cramér ([1]). Another set of conditions is given in Dacunha-Castelle and Gassiat ([2]), also. Then, the Taylor expansion leads us to

$$L_n(\alpha, \hat{\beta}_\alpha) = \frac{1}{2n} U_n(\alpha)^2 + o_p(1/n), \tag{27}$$

where $U_n(\alpha)$ is the empirical process defined by

$$U_n(\alpha) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n v_\alpha(Z_i)}{\sqrt{\frac{1}{n} \sum_{i=1}^n v_\alpha(Z_i)^2}}, \tag{28}$$

and $v_\alpha(z)$ is a function in the basis of the tangent cone defined by

$$v_\alpha(z) = \frac{\partial}{\partial\beta} \log f(z; (\alpha, 0)). \tag{29}$$

The denominator of $U_n(\alpha)$ converges to one and the numerator converges in law to the standard normal distribution for each $\alpha \in A_0$. If we consider the behavior of $U_n(\alpha)$ over all α, it can be looked as a stochastic process over α or C, and all the marginal distributions converges to a multidimensional normal distribution. If the higher order term of $o_p(1/n)$ is bounded uniformly over α, and the stochastic process U_n converges uniformly to a Gaussian process, the limit of the supremum of $nL_n(\alpha, \hat{\beta}_\alpha)$ over α can be replaced by the square of the supremum of the Gaussian process. Dacunha-Castelle and Gassiat ([2]) discussed this case, assuming that the function class $C = \{v_\alpha(z)\}$ is Donsker.

Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(\mathcal{Z}, \mathcal{B})$ be a measurable space, and $Z_1, Z_2, \ldots$ be i.i.d. random variables with their value in $\mathcal{Z}$. A family of Borel measurable functions $\mathcal{F} \subset \{v : \mathcal{Z} \to \mathbb{R}\}$ is called *Donsker* if $E_P[v(Z)]$

and $E_P[v(Z)^2]$ exist for all $v \in \mathcal{F}$, the map $z \mapsto \sup_{v\in\mathcal{F}} |v(z)|$ is finite for every $z \in \mathcal{Z}$, and the $\mathcal{F}$-indexed empirical processes

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (v(Z_i) - E_P[v(Z)]), \tag{30}$$

as considered to be random elements with their values in the Banach space $\ell^\infty(\mathcal{F})$ of all the bounded functions on $\mathcal{F}$ with sup norm, converge in law to a tight[2] Borel measurable random element with its value in $\ell^\infty(\mathcal{F})$.

In discussing the stochastic process in eq.(27), we will investigate both of Donsker and non-Donsker cases. For Donsker cases, Dacunha-Castelle and Gassiat ([2]) clarify the limiting distribution of likelihood ratio of the maximum likelihood estimator, and apply the result to finite mixture models and ARMA models. In this paper, we will derive a relation between the likelihood ratio and the Kullback-Leibler divergence of the maximum likelihood estimator in Donsker cases, as a simple consequence of their result. In non-Donsker cases, a diversity of phenomena can be seen. Even the order of the likelihood ratio may be different from the usual $O_p(1/n)$. Hartigan ([3]) reported a larger order than $O_p(1/n)$ in the likelihood ratio test of the normal mixture model with two components. Hagiwara et al. ([4]) elucidated the order $O(\log n/n)$ for the likelihood ratio of the maximum likelihood estimator in multilayer perceptron models. We will derive a useful sufficient condition of such a larger order than $O_p(1/n)$ of the likelihood ratio of the maximum likelihood estimator, in terms of the functional property of the basis of the tangent cone.

3.2 Donsker cases

To apply the theory of convergence to a Gaussian process, we have to assure the uniformity over α of the small order in eq.(27). First, for the uniform consistency of $\hat{\beta}_\alpha$, we need the following uniform Wald conditions.

[Uniform Wald conditions (W)]

1. There exists a set E with $f(z)\mu$-probability 1 such that for any z in E and any α,

$$\lim_{|\beta|\to\infty} f(z; (\alpha, \beta)) = 0. \tag{31}$$

[2]Let $\mathcal{X}$ be a topological space, and $(\mathcal{X}, \mathfrak{S})$ be the Borel measurable space. A Borel measurable random variable $Z : \Omega \to \mathcal{X}$ is called *tight* if for arbitrary ε there exist a compact set K in $\mathcal{X}$ such that $P(Z \in K) \geq 1 - \varepsilon$.

2. Consider the functions

$$F(z;\beta,\rho) := \sup_{\substack{|\beta'-\beta|\le\rho \\ \alpha}} f(z;\beta',\alpha), \qquad G(z;r) := \sup_{\substack{|\beta|\ge r \\ \alpha}} f(z;\beta',\alpha) \quad (32)$$

for $\rho > 0$ and $r > 0$, and define $F^*(z;\beta,\rho) = \max\{F(z;\beta,\rho), 1\}$ and $G^*(z;\rho) = \max\{G(z;\rho), 1\}$. Then, the following conditions hold;

$$\lim_{\rho\to+0} E_{f_0(z)\mu}[\log F^*(z;\beta,\rho)] < \infty, \qquad \lim_{r\to\infty} E_{f_0(z)\mu}[\log G^*(z;r)] < \infty. \quad (33)$$

Using the same discussion in Wald ([14]), under the above conditions (W), the maximum likelihood estimator in the submodel $\hat{\beta}_\alpha$ converges to 0 in probability uniformly over α.

To assure the uniformly small order of $o_p(1/n)$, we further assume the following condition:

[Uniformity condition (U)]

Consider the functions

$$H_1(z;\beta,\rho) := \sup_{\substack{|\beta'-\beta|\le\rho \\ \alpha}} \left| \frac{\frac{\partial}{\partial\beta} f(z;\beta',\alpha)}{f(z;\beta',\alpha)} \right|, \quad K_1(z;r) := \sup_{\substack{|\beta|\ge r \\ \alpha}} \left| \frac{\frac{\partial}{\partial\beta} f(z;\beta',\alpha)}{f(z;\beta',\alpha)} \right|,$$

$$H_2(z;\beta,\rho) := \sup_{\substack{|\beta'-\beta|\le\rho \\ \alpha}} \left| \frac{\frac{\partial^2}{\partial\beta^2} f(z;\beta',\alpha)}{f(z;\beta',\alpha)} \right|, \quad K_2(z;r) := \sup_{\substack{|\beta|\ge r \\ \alpha}} \left| \frac{\frac{\partial^2}{\partial\beta^2} f(z;\beta',\alpha)}{f(z;\beta',\alpha)} \right|. \quad (34)$$

Then, the following conditions hold for $i = 1, 2$;

$$\lim_{\rho\to+0} E_{f_0(z)\mu}[(H_i(z;\beta,\rho))^2] < \infty, \qquad \lim_{r\to\infty} E_{f_0(z)\mu}[K_i(z;r)] < \infty. \quad (35)$$

The following theorem is due to Dacunha-Castelle and Gassiat ([2]).

Theorem 1. *Let a statistical model $S = \{f(z;(\alpha,\beta)\}$ be locally conic at $f_0(z)$. Assume (W) and (U) hold, and the family of functions $C = \{v_\alpha(z) = \frac{\partial}{\partial\beta} f(z;(\alpha,0)\}$ is Donsker. then the supremum of the likelihood ratio converges in law as follows;*

$$n \sup_{(\alpha,\beta)} L_n(\alpha,\beta) \longrightarrow \frac{1}{2} \sup_{v\in C} W^2, \quad (36)$$

where W is a tight, Borel measurable Gaussian process over C, which is a limit of the empirical process U_n.

A sufficient condition of the Donsker is known ([15]). A class of functions $\mathcal{F}$ is Donsker if (i) the envelop function $F(z) = \sup_{v \in \mathcal{F}} |v(z)|$ is P-(outer) square integrable, (ii) the square root of the uniform entropy number is integrable, and (iii) P-measurability on some function classes are satisfied.

In these three conditions, the measurability conditions are automatically satisfied if $\mathcal{F} = \{w(z;a)\}$ is parameterized by a separable metric space and $w(z;a)$ is continuous about a for all z. This is true for the basis of the tangent cone of a locally conic model. A sufficient condition for integrability of the uniform entropy number is that the VC-dimension of $\mathcal{F}$ is finite. These are often satisfied by the tangent cone of many models, such as neural networks.

Note that the condition (i) is satisfied if the integral of the square of $H_1(z;0,\rho)$ is finite for a sufficiently small ρ. Therefore, we obtain the following corollary.

Corollary 1. *Let a statistical model $S = \{f(z;(\alpha,\beta)\}$ be locally conic at $f_0(z)$. Assume (W) and (U) hold, and the VC-dimension of $C = \{v_\alpha(z) = \frac{\partial}{\partial\beta} f(z;(\alpha,0)\}$ is finite. Then, C is Donsker, and eq.(36) holds for a tight, Borel measurable Gaussian process W.*

In Donsker cases, we can derive a simple relation between the likelihood ratio and the Kullback-Leibler divergence, which is satisfied by regular models.

Theorem 2. *Under the same assumptions as Theorem 1 or Corollary 1, D and L_n have the order of $O_p(1/n)$, and the relation*

$$D(\hat{\alpha},\hat{\beta}) = L_n(\hat{\alpha},\hat{\beta}) + o_p(1/n) \tag{37}$$

holds.

Proof. The standard argument of Taylor expansion of D with respect to β gives the second argument. Since W is a tight Gaussian process, the class C is necessarily totally bounded in $L^2(P)$, and almost all the sample paths $v \mapsto W(v)$ are uniformly $L^2(P)$ continuous (see van der Vaart and Wellner [15], Section 1.5). Then, the supremum of $|W|$ is finite almost surely. □

The above result holds also to a regular model, which satisfies the asymptotic efficiency. We can not obtain the exact distribution of the likelihood ratio or Kullback-Leibler divergence in non-regular cases. In non-Donsker cases, a clear relation as eq.(37) has not been known.

3.3 Non-Donsker cases

As we mentioned in Section 3.1, the likelihood ratio of the maximum likelihood estimator does not necessarily have the usual order $O_p(1/n)$, but can have a larger order, if the function class of the tangent cone is "rich" enough like normal mixtures and multilayer perceptrons.

We derive a useful sufficient condition of such an unusually larger order, extending Hartigan's idea. Note that a marginal of U_n on finitely many points $v_1, \ldots, v_m$ in C always converges to a multi-dimensional normal distribution. The covariance of the limit is given by

$$E_P[v_i v_j]. \tag{38}$$

The two components are independent if their covariance is zero. Suppose we can find arbitrary number of "almost" independent Gaussian random variables in C, then, the supremum of $U_n(\alpha)$ on such variables can take an arbitrary large value, since the maximum of m independent samples from the standard normal distribution is $\sqrt{2\log m}$ for infinitely large m. Hartigan ([3]) applied this idea to a normal mixture model with two components, calculating the covariance explicitly. An extension of this idea leads us to the following theorem;

Theorem 3. *Let a statistical model $S = \{f(z;(\alpha,\beta))\}$ be locally conic at $f_0(z)$, and $C = \{v_\alpha(z) = \frac{\partial}{\partial\beta} f(z;(\alpha,0))\}$ be the basis of the tangent cone. Suppose there exists a sequence $\{v_n\}_{n=1}^{\infty}$ in C such that $v_n(z) \to 0$ almost every z, then, for arbitrary $M > 0$*

$$\lim_{n\to\infty} \Pr\Big(\sup_{(\alpha,\beta)} nL_n(\alpha,\beta) \le M\Big) = 0 \tag{39}$$

Proof. From Proposition 2 below, for arbitrary $\varepsilon > 0$ and $K \in \mathbb{N}$, there exist $v(\alpha_1), \ldots, v(\alpha_K) \in C$ such that $|E[v(\alpha_i)v(\alpha_j)]| < \varepsilon$ for different i and j. Then, the rest of the proof is the same as Hartigan ([3]).

Let Σ be the variance-covariance matrix of the K-dimensional normal distribution, which is the limit of the empirical process over $v(\alpha_1), \ldots, v(\alpha_K)$, and $W = (W_1, \ldots, W_K)$ be a random vector following the limiting normal distribution. Because of the fact $-\frac{1}{2}W^T\Sigma^{-1}W \le -\frac{1}{2}W^TW + \varepsilon\sum_{i<j} W_iW_j \le$

$-\frac{1-\varepsilon}{2}W^TW$, we have for arbitrary $M > 0$

$$\begin{aligned} P\big(\max_{1\le i\le K}|W_i| \le M\big) &\le \int_{[-M,M]^K} \frac{1}{\sqrt{(2\pi)^K|\Sigma|}} e^{-\frac{1-\varepsilon}{2}w^2} dw \\ &= \frac{1}{|\Sigma|^{K/2}(1-\varepsilon)^{K/2}} \int_{[-(1-\varepsilon)M,(1-\varepsilon)M]^K} \frac{1}{(2\pi)^{K/2}} e^{-\frac{1}{2}u^Tu} du \\ &\le \frac{1}{(1-K\varepsilon)^K(1-\varepsilon)^{K/2}} (\Phi(M) - \Phi(-M))^K, \end{aligned} \tag{40}$$

where $\Phi(t)$ is the cumulative distribution function of the standard normal distribution. Taking ε so that $\frac{1}{(1-K\varepsilon)^K(1-\varepsilon)^{K/2}} < 2$, we have the inequality

$$P\big(\max_{1\le i\le K}|W_i| \le M\big) < 2(\Phi(M) - \Phi(-M))^K. \tag{41}$$

The convergence of $(U_n(\alpha_1), \dots, U_n(\alpha_K))$ to W means $\lim_{n\to\infty} P(\max_i |U_n(\alpha_i)| \le M) = P(W \in [-M, M]^K)$. Therefore, we obtain

$$\lim_{n\to\infty} P\big(\max_{1\le i\le K}|U_n(\alpha_i)| \le M\big) \le 2(\Phi(M) - \Phi(-M))^K. \tag{42}$$

Since $(\Phi(M)-\Phi(-M))^K$ takes an arbitrary small value for sufficiently large K, the assertion is proved. □

On the covariance of the random variables with bounded L^2 norm, we have the following proposition.

Proposition 2. *Let $\{v_n\}_{n=1}\infty$ be a sequence in $L^2(P)$ such that $\|v_n\|_{L^2(P)} = 1$ for all n, and $v_n(x) \to 0$ for almost every x. Then, for all n and $\varepsilon > 0$, there exists ℓ_0 such that*

$$E_P|v_nv_\ell| < \varepsilon \tag{43}$$

for all $\ell \ge \ell_0$.

This is a direct consequence of the following proposition.

Proposition 3. *Let $(\Omega, \mathcal{B}, P)$ be a probability space, and $Y, X_1, X_2, \dots$ be random variables. Suppose that $\int Y^2 dP$ and $\int X_n^2 dP$ are upper bounded by K, and X_n converges to 0 in probability. Then, we have*

$$\lim_{n\to\infty} E|YX_n| = 0. \tag{44}$$

Proof. Let ε be an arbitrary positive number. Because $\int Y^2 dP < \infty$, there exists $\delta > 0$ such that $\int_\Delta Y^2 dP < \frac{\varepsilon^2}{9K}$ for any measurable set Δ with $P(\Delta) < \delta$.

For each n, define a set

$$A_n = \{\omega \in \Omega \mid |Y| > \frac{\varepsilon}{3\sqrt{K}} \quad \text{and} \quad |X_n| > \frac{\varepsilon}{3K}|Y|\}. \tag{45}$$

Because $X_n \to 0$ in probability and $A_n \subset \{|X_n| > \frac{\varepsilon^2}{9K^{3/2}}\}$, we can find n_0 such that $P(A_n) < \delta$ for all $n \geq n_0$. Then, we have $\int_{A_n} Y^2 dP < \frac{\varepsilon^2}{9K}$ for $n \geq n_0$.

For $n \geq n_0$, we derive

$$\begin{aligned}
&\int |YX_n| dP = \int_{A_n} |YX_n| dP + \int_{A_n^c} |YX_n| dP \\
&\leq \Big(\int_{A_n} Y^2 dP\Big)^{1/2} \Big(\int_{A_n} X_n^2 dP\Big)^{1/2} + \int_{\{|Y| \leq \frac{\varepsilon}{3\sqrt{K}}\}} |YX_n| dP + \int_{\{|X_n| \leq \frac{\varepsilon}{3K}|Y|\}} |YX_n| dP \\
&\leq \frac{\varepsilon}{3\sqrt{K}}\sqrt{K} + \frac{\varepsilon}{3\sqrt{K}} \int |X_n| dP + \frac{\varepsilon}{3K} \int |Y|^2 dP \\
&\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3\sqrt{K}} \cdot \sqrt{K} + \frac{\varepsilon}{3K} \cdot K = \varepsilon
\end{aligned} \tag{46}$$

In the last line, we use the fact $\int |X_n| dP \leq (\int |X_n|^2 dP)^{1/2} \leq \sqrt{K}$. □

3.4 Likelihood of multilayer perceptrons

We apply the results in the previous subsections to multilayer perceptrons. For simplicity, we discuss networks without bias terms in the output unit:

$$\varphi(x;\theta) = \sum_{j=1}^{H} b_j \, s(a_j x + c_j), \tag{47}$$

and assume the constant zero true function $\varphi_0(x) = 0$. In this case, we can introduce a locally conic parameterization, by taking a parameter set $A = \{(a_1, \ldots, a_H, c_1, \ldots, c_H, \eta_1, \ldots, \eta_H) \in \mathbb{R} \mid 0 < a_1 < \cdots < a_H, \quad \sum_{j=1}^H \eta_j^2 = 1, \quad \eta_j \neq 0 \ (1 \leq j \leq H) \quad \eta_1 > 0\}$, and defining

$$\psi(z;(\alpha,\beta)) = \beta \sum_{j=1}^{H} \eta_j \, s(a_j x + c_j) \tag{48}$$

for $(\alpha, \beta) \in A \times \mathbb{R}$.

We assume the exponential family for the noise model $r(y|u)$, which is discussed in Section 2.5. Without loss of generality, we assume that the variance of the y is one, that is, $\int y^2 r(y|0)dy = 1$. The basis of the tangent cone C is given by

$$v_\alpha(x, y) = y \frac{\sum_{j=1}^H \eta_j s(a_j x + c_j)}{\| \sum_{j=1}^H \eta_j s(a_j x + c_j)\|_{L^2(Q)}}. \tag{49}$$

It is easy to see that C includes a sequence converging to constant zero almost everywhere, if $H \geq 2$. In fact, we can find such a sequence by $c \to 0$ and $a_1, a_2 \to \infty$ for $y\{s(a_1 x + c) - s(a_2 x - c)\}/\|s(a_1 x + c) - s(a_2 x - c)\|_{L^2(Q)}$. Therefore, we obtain the following

Theorem 4. *Assume the true function is constant zero, and the noise model $r(y|u)$ is an exponential family $r(y|u) = \exp(y\kappa(u) + \tau(y) - \zeta(u))$ satisfying $\kappa'(u) \neq 0$ and $\int y r(y|u)dy = u$. For the multilayer perceptron model (47) with more than one hidden unit, we have*

$$\lim_{n\to\infty} \Pr\Big(\sup_{(\alpha,\beta)} nL_n(\alpha, \beta) \leq M \Big) = 0. \tag{50}$$

We can derive a tighter lower bound of the likelihood ratio in the above problem, by counting a number of almost independent random variables in C.

Theorem 5. *Under the same assumptions as Theorem 4, we have*

$$\sup_\theta L_n(\theta) \gtrsim O_p\big(\tfrac{\log n}{n}\big), \tag{51}$$

as n goes to infinity.

Rough sketch of the proof. Take the function in C defined by

$$y \frac{s(a_1(x - (c + \delta))) - s(a_2(x - (c - \delta)))}{\|s(a_1(x - (c + \delta))) - s(a_2(x - (c - \delta)))\|_{L^2(Q)}}. \tag{52}$$

For sufficiently large a_1, a_2, and sufficiently small $\delta > 0$, we obtain a random variable with its mass concentrated in a arbitrary small interval around c. Since c is arbitrary, we can obtain any number of random variables in C with their covariance arbitrary small. We can prepare n^γ ($\gamma > 0$) number of such variables assuring the uniform convergence of empirical process $U_n(v)$

to a Gaussian distribution. From the extreme value theory, the supremum of m i.i.d. samples from the chi-square distribution is $2\log m$. Using the fact that the n^γ random variables are arbitrary close to be independent, we can prove that the supremum of $|U_n(v)|^2$ over the variables is of the order $\log n^\gamma = \gamma \log n$. □

The order $O_p(\log n/n)$ has been formerly obtained by Hagiwara et al. ([4]). However, they assume the additive Gaussian noise model. Our approach extends their results. The above theorem can be applied to various noise models, including binary output models.

As we can see in the above discussions, the behavior of the likelihood ratio deeply depends on the functional property of the tangent cone C. If the multilayer perceptron model has only one hidden unit, the behavior is totally different. In this case, the basis of the tangent cone is Donsker, and we can apply Theorem 2. obtain

Theorem 6. *Under the same assumptions as Theorem 4, for the multilayer perceptron model (47) with one hidden unit, we have*

$$D(\hat{\theta}) = L_n(\hat{\theta}) + o_p(1/n), \quad \textit{and} \quad L_n(\hat{\theta}) = O_p(1/n). \tag{53}$$

We omit the proof. □

4 Conclusion

We have discussed an approach to investigate the behavior of the maximum likelihood estimator in the case that the true parameter is not identifiable. We have seen that the unidentifiability of parameters in a statistical model can be formulated by a conic singularity in many cases. Following the discussion of Dacunha-Castelle and Gassiat ([2]), we have formulated the likelihood ratio of the maximum likelihood estimator by the supremum of an empirical process, which converges to the standard normal distribution marginally. Rather than concentrating on Donsker cases, we have discussed non-Donsker cases, and derived a useful sufficient condition of an unusual larger order of the likelihood ratio. We have applied these results on neural network models, and derived the lower bound of the likelihood ratio, assuming that the true function is constant zero.

The omitted proofs will be presented in a forthcoming paper.

Acknowledgements

I thank Prof. Kano in Osaka University for teaching me Hartigan's paper, and Prof. Kuriki in the Institute of Statistical Mathematics for valuable discussions.

References

[1] Cramér, H. (1946) *Mathematical Methods of Statistics*. Princeton University Press.

[2] Dacunha-Castelle, D. and Gassiat, E. (1997) Testing in locally conic models, and application to mixture models *ESAIM Probability and Statistics*, **1**, 285–317.

[3] Hartigan, J.A. (1985). A failure of likelihood asymptotics for normal mixtures. *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer*, vol.II, pp.807–810.

[4] Hagiwara, K., Kuno K., and Usui S. (2000) On the problem in model selection of neural network regression in overrealizable scenario. *Proceeding of International Joint Conference of Neural Networks*.

[5] Fukumizu, K. (1999) Generalization error of linear neural networks in unidentifiable cases. O.Watanabe and T.Yokomori (eds.) Lecture Notes in Artificial Intelligence 1720, Algorithmic Learning Theory (Proceedings of the 10th International Conference on Algorithmic Learning Theory (ALT'99)), pp.51-62. Springer-Verlag: Berlin.

[6] Csörgö, M. and Horváth, L. (1996) *Limit Theorems in Change-Point Analysis*. John Wiley & Sons.

[7] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) in: Learning internal representations by error propagation, eds. D.E. Rumelhart, J.L. McClelland and the PDP Research Group, Parallel distributed processing, Vol.1 (MIT Press, Cambridge) pp.318–362.

[8] Pistone, G. and Sempi, C. (1995). An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The Annals of Statistics*, 23(5):1543–1561.

[9] Sussmann, H.J. (1992) Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, **5**, 589–593.

[10] Chen, A. M., Lu, H., and Hecht-Nielsen, R. (1993). On the geometry of feedforward neural network error surfaces. *Neural Computation*, **5**, 910–927.

[11] Fukumizu, K. and Amari, S. (2000) Local Minima and Plateaus in Hierarchical Structures of Multilayer Perceptrons. *Neural Networks*, **13**(3), 317–327.

[12] Fukumizu, K. (1996) A regularity condition of the information matrix of a multilayer perceptron network. *Neural Networks*, **9**(5), 871–879.

[13] Sen, P.K. and Singer, J.M. (1993) *Large sample methods in statistics.* Chapman & Hall.

[14] Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20**, 595–601.

[15] Van der Vaart, A.W. & Wellner, J.A. (1996). *Weak convergence and empirical processes.* Springer Verlag.