Journal of the Korean Statistical Society (2001), **30**: *2, pp 179–192*

Statistical Inference in Non-Identifiable and Singular Statistical Models

Shun-ichi Amari[1], Hyeyoung Park[1], and Tomoko Ozeki[1]

ABSTRACT

When a statistical model has a hierarchical structure such as multilayer perceptrons in neural networks or Gaussian mixture density representation, the model includes distributions with unidentifiable parameters when the structure becomes redundant. Since the exact structure is unknown, we need to carry out statistical estimation or learning of parameters in such a model. From the geometrical point of view, distributions specified by unidentifiable parameters become a singular point in the parameter space. The problem has been remarked in many statistical models, and strange behaviors of the likelihood ratio statistics, when the null hypothesis is at a singular point, have been analyzed so far.

The present paper studies asymptotic behaviors of the maximum likelihood estimator and the Bayesian predictive estimator, by using a simple cone model, and show that they are completely different from regular statistical models where the Cramér-Rao paradigm holds. At singularities, the Fisher information metric degenerates, implying that the Cramér-Rao paradigm does no more hold, and that the classical model selection theory such as AIC and MDL cannot be applied. This paper is a first step to establish a new theory for analyzing the accuracy of estimation or learning at around singularities.

Keywords: Singular structure; Hierarchical system; Maximum Likelihood Estimator; Bayesian Predictive Distribution

1. INTRODUCTION

It has been known that some statistical models such as Gaussian mixtures and changing point estimation include unidentifiable parameters in its part. Hierarchical systems such as neural networks also include complex singular structures in the parameter spaces. A lower-order system is included in the space of a higher-order system as a subset and when a true parameter is on such a subspace, the

[1]Brain Science Institute, RIKEN, Wako, Saitama, 351-0198, Japan

true parameter is unidentifiable. Here, the Fisher information matrix degenerates, and the conventional paradigm of the Cramér-Rao bound does not hold. However, we need to estimate the parameters of such a system by learning, which is a type of sequential estimation, in many engineering problems using artificial neural networks (Amari, 1998; Amari, Park, and Fukumizu, 2000; Park, Amari, and Fukumizu, 2000). Therefore, it is important to establish a different paradigm for analyzing various characteristics of learning or estimation in singular models including points of unidentifiable parameters.

There have been a number of studies on this problem of singularity or unidentifiability. In many statistical literatures, strange behaviors of the log likelihood ratio have been studied, when the null hypothesis is at an unidentifiable or singular point. Hagiwara et al. (2000) investigated abnormal phenomena at the singularities by using neural networks models and pointed out that the AIC type criterion of model selection is no more valid. Fukumizu (2000; 2001) analyzed the behaviors of the maximum likelihood estimator in unidentifiable situations by applying the ideas of Hartigan (1985) and Dacunha-Castelle and Gassiat (1997). Watanabe (2001a; 2001b) also analyzed the behaviors of the Bayesian predictive distribution in the algebraic-geometrical framework. Amari and Ozeki (2001) showed the influence of the singularities on the behavior of learning (Amari, Park, and Fukumizu, 2000; Park, Amari, and Fukumizu, 2000) by using a simple toy model. However, these studies have been conducted separately, from different viewpoints. We need a more integrative framework to investigate the whole characteristics of this type of statistical inference under models including singularities. This is also an important subject of research in information geometry (Amari and Nagaoka, 2000).

The generalization error is a basic factor for investigating the characteristics of a stochastic model. It represents how well an estimated system behaves. It is given by the Kullback-Leibler divergence of the estimated distribution from the unknown true distribution. Model selection as well as parameter estimation intends to minimize the generalization error. However, the generalization error cannot be directly evaluated, and we use the training error, which is the empirical loss and computable. To this end, we need to evaluate the bias of the training error, which leads us to the AIC criterion for model selection. Indeed, for regular statistical models, the gap between the generalization error and the training error is given by a term, depending only on the number of the parameters divided by the number of examples. This fact has been shown by Amari and Murata (1993) to hold in general neural network models and by many others. This result is the

base of the well known criterion, AIC, for model selection. However, when one compares two hierarchical models, the lower-order model includes unidentifiable parameters in the space of the higher-order model. Therefore, the conventional results mentioned above cannot be applied, and we need a new result estimating the gap between the generalization error and the training error at singularities.

The present paper analyzes the generalization error and training error in the framework of the Gaussian random field, used by Dacunha-Castelle and Gassiat (1997), Fukumizu (2000), and Hartigan (1985). We use a simple cone model to investigate the relationship between the generalization error and the training error for the maximum likelihood estimator and the Bayesian predictive distribution. The present results show how the asymptotic behaviors of estimators differ from the regular case, elucidating the strange behaviors of singular models in non-Cramér-Rao paradigm.

In the next section, we summarize known asymptotic results for parameter estimation. We then define the problem including singularities (unidentifiabilities) in the parameter space of probability density functions, and give some typical examples of statistical models with singularities. In section 4, we describe our method to analyze the generalization error and the training error of MLE, and give some results for a simple toy model. In section 5, we describe the method of analysis for the Bayesian predictive distributions, and give interesting new results. The conclusions and discussions for future works are given in section 6.

2. ASYMPTOTICS OF STATISTICAL ESTIMATION

Let us begin with a statistical model $S = \{p(\boldsymbol{x}|\boldsymbol{\theta})|\boldsymbol{\theta} \in \Theta\}$, which is defined by a probability density function of a random variable $\boldsymbol{x}$. The probability density function is specified by the parameter $\boldsymbol{\theta}$ in the parameter space Θ. We assume that the true probability density $p_o(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta}_o)$ is in the model S. When a sample of observations $D = \{\boldsymbol{x}_1, \ldots \boldsymbol{x}_n\}$ generated by the true probability density $p_o(\boldsymbol{x})$ is given, we try to find an estimated distribution $\hat{p}(\boldsymbol{x})$, which is a plug-in distribution $p(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$ of estimator $\hat{\boldsymbol{\theta}}$ or the Bayes predictive estimator, through minimizing the distance between $p(\boldsymbol{x})$ and $p_o(\boldsymbol{x})$. The distance between two probability density functions is measured by the Kullback-Leibler divergence of the form,

$$K[p_o : \hat{p}] = E_{p_o}\left[\log \frac{p_o(\boldsymbol{x})}{\hat{p}(\boldsymbol{x})}\right]. \tag{2.1}$$

The expectation of the negative $\log \hat{p}(\boldsymbol{x})$, $E_{p_o}[-\log \hat{p}(\boldsymbol{x})]$, is sometimes called the generalization error, and is given by

$$E_{gen} = H_o + E_D\left[K[p_o : \hat{p}]\right], \tag{2.2}$$

where H_o is the entropy of $p_o(\boldsymbol{x})$ and E_D denotes expectation with respect to observed data. Similarly, the training error is defined by using the empirical expectation,

$$E_{train} = H_o + E_D\left[\frac{1}{n}\sum_{i=1}^{n} \log \frac{p_o(\boldsymbol{x}_i)}{\hat{p}(\boldsymbol{x}_i)}\right]. \tag{2.3}$$

These are terminologies from neural networks community. In order to evaluate the estimator $\hat{p}$, one uses E_{gen} or $E_D\left[K[p_o : \hat{p}]\right]$, but it is not computable. Instead, one uses the arithmetic mean of $\log[p_o(\boldsymbol{x}_i)/\hat{p}(\boldsymbol{x}_i)]$ of the data whose expectation gives $E_{train} - H_o$, which is computable. Hence, it is important to see the difference between E_{gen} and E_{train}. This is the principle of AIC in model selection.

When the statistical model S is regular, or the true distribution $p_o(\boldsymbol{x})$ is identifiable, the mle-based $p(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ and the Bayes predictive distribution are known to be Fisher efficient under reasonable regularity conditions,

$$E_D\left[K[p_o : \hat{p}]\right] \approx \frac{d}{2n}, \tag{2.4}$$

where d is the dimension number of parameters $\boldsymbol{\theta}$. It is also proved that

$$E_{gen} \approx E_{train} + \frac{d}{2n} \tag{2.5}$$

asymptotically. AIC is a criterion to estimate E_{gen} from the above relation for model selection.

We will show most of these good relations do not hold when $p_o(\boldsymbol{x})$ is unidentifiable.

3. MODELS WITH SINGULARITIES

In a statistical model $S = \{p(\boldsymbol{x}|\boldsymbol{\theta})|\boldsymbol{\theta}\}$, when $p(\boldsymbol{x}|\boldsymbol{\theta}_1) = p(\boldsymbol{x}|\boldsymbol{\theta}_2)$ holds for $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$, the two points $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are said to be equivalent. When the set of equivalent points forms a submanifold in the parameter space, the Fisher information matrix degenerates on it, and the parameters are unidentifiable when the true one is on the submanifold. Dividing the parameter space by the equivalent

relation, all equivalent points reduce to one class. This causes singularities in the reduced space such that dimensions are reduced in such unidentifiable sets. These singularities are very ubiquitous in the space of hierarchical statistical models such as neural networks and Gaussian mixture models.

In the present paper, we discuss singularities with a cone structure. Let us first divide the parameter $\boldsymbol{\theta}$ into two parts; ξ and $\boldsymbol{\omega}$, $\boldsymbol{\theta} = (\xi, \boldsymbol{\omega})$, and assume that the probability density for the statistical model is represented by

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = p(\boldsymbol{x}|\xi\phi(\boldsymbol{\omega})). \tag{3.1}$$

For this type of models, the parameter $\boldsymbol{\omega}$ is unidentifiable when $\xi = 0$. This cone structure of singularity occurs in various statistical models including unidentifiable parameters (Dacunha-Castelle and Gassiat, 1997). In this section, we introduce some examples of statistical models that have such a singularity.

3.1. Cone Model

We first consider the set of Gaussian distributions of random variable $\boldsymbol{x} \in R^{d+2}$, with mean $\boldsymbol{\mu}$ and identity covariance matrix I,

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \frac{1}{(\sqrt{2\pi})^{d+2}} \exp\left\{-\frac{1}{2}|\boldsymbol{x} - \boldsymbol{\mu}|^2\right\} \tag{3.2}$$

This is the enveloping model S. The cone model M is a subset of S, embedded as

$$M : \boldsymbol{\mu} = \frac{\xi}{\sqrt{1+c^2}} \begin{pmatrix} 1 \\ c\boldsymbol{\omega} \end{pmatrix} = \xi \boldsymbol{a}(\boldsymbol{\omega}) \tag{3.3}$$

$$\boldsymbol{\omega} \in S^d, \quad |\boldsymbol{a}|^2 = 1, \tag{3.4}$$

where c is a constant and S^d is a d-dimensional unit sphere. When $d = 1$, S^1 is a circle so that $\boldsymbol{\omega}$ is replaced by angle θ, and we have

$$\boldsymbol{\mu} = \frac{\xi}{\sqrt{1+c^2}} \begin{pmatrix} 1 \\ c\cos\theta \\ c\sin\theta \end{pmatrix}. \tag{3.5}$$

See Figure 3.1. The M is a cone, having $(\xi, \boldsymbol{\omega})$ as coordinates, where the apex $\xi = 0$ is a singular point with the cone structure. From the next section, we will mainly discuss this model for analyzing the generalization error and the training error.

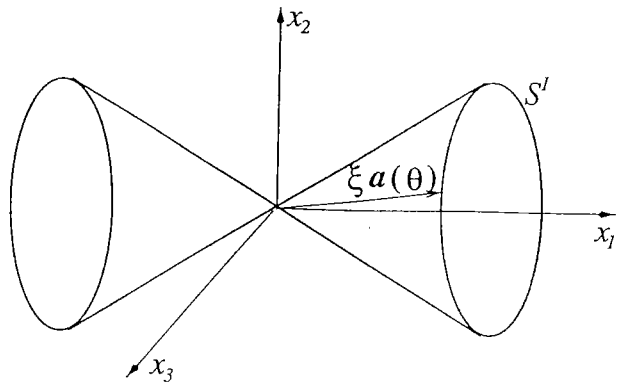

Figure 3.1: One-dimensional Cone Model

3.2. Neural Networks

A multilayer perceptron, which receives an input vector signal $\boldsymbol{x}$ and emit a scalar output signal y, can also be considered as a stochastic model. Let h be the number of hidden units, and let $\boldsymbol{w}_i$ be the weight vector of the ith hidden unit, $i = 1, \cdots, h$. Let φ be the sigmoidal activation function such as hyperbolic tangent, and let v_i be the weight from the ith hidden unit to the output unit. We assume that the output unit is linear, but is disturbed by Gaussian noise n with mean 0 and variance σ^2. The input-output relation of a multilayer perceptron is then represented as

$$y = f(\boldsymbol{x}, \boldsymbol{\theta}) + n \tag{3.6}$$

where

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{h} v_i \varphi\left(\boldsymbol{w}_i \cdot \boldsymbol{x}\right), \tag{3.7}$$

and $\boldsymbol{\theta} = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_h, v_1, \cdots, v_h)$ denote modifiable parameters. Because of the noise n, the behavior of a system is described by the conditional probability density function of output y conditioned on input $\boldsymbol{x}$,

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\left\{y - f(x, \theta)\right\}^2\right\}. \tag{3.8}$$

Let us denote by $M(h)$ the set of all the perceptrons with h hidden neurons, and identify each point of the space $M(h)$ with the associated (conditional) probability distribution. In other words, $M(h)$ is regarded as a statistical model consisting of such probability distributions parameterized by $\boldsymbol{\theta}$.

The model $M(h)$ has a hierarchical structure, since it includes $M(h-1), M(h-2), \ldots$ as subspaces. For example, when

$$v_i \boldsymbol{w}_i = 0 \tag{3.9}$$

holds, the ith hidden neuron does not play any role so that it can be removed. Hence, the subspace defined by (3.9) corresponds to $M(h-1)$. When $\boldsymbol{w}_i = \boldsymbol{w}_j$, the ith and jth neurons play the same role so that they can be merged into one neuron. Hence, the subspace given by $\boldsymbol{w}_i = \boldsymbol{w}_j$ is also identified with $M(h-1)$. The parameters of these subspaces make singularities in the space of $M(h)$. For the simplest case, let us consider the model $M(1)$ of only one hidden neuron, which has the function mapping of the form,

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = v\varphi(\boldsymbol{w} \cdot \boldsymbol{x}). \tag{3.10}$$

Then we can find the singularity of cone structure at $v = 0$.

3.3. Gaussian Mixtures

The Gaussian mixture is a weighted sum of Gaussian probability density functions,

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{h} v_i \psi(\boldsymbol{x} - \boldsymbol{\mu}_i), \quad \sum v_i = 1, \tag{3.11}$$

where ψ is the Gaussian density function,

$$\psi(\boldsymbol{x}) = c \exp\left\{-\frac{1}{2}|\boldsymbol{x}|^2\right\}. \tag{3.12}$$

The set of the Gaussian mixtures with h components forms a space $M(h)$, which includes $M(h-1)$ as a subspace. Hence, the Gaussian mixtures is also a kind of hierarchical model with complex singularities. In addition, we can find that it has the cone structure of singularity when $v_i = 0$ or $\boldsymbol{\mu}_i = \boldsymbol{\mu}_j$.

4. GENERALIZATION ERROR AND TRAINING ERROR OF MAXIMUM LIKELIHOOD ESTIMATOR

For the sake of simplicity, we use the simple cone model in order to analyze the generalization error and the training error of the mle. However, our method is applicable to other models such as neural networks. For a given set of

observations, $D = \{\boldsymbol{x}_i\}_{i=1,\ldots,n}$, the log likelihood of D is written as

$$L(D, \xi, \boldsymbol{\omega}) = -\frac{1}{2}\sum_{i=1}^{n} |\boldsymbol{x}_i - \xi \boldsymbol{a}(\boldsymbol{\omega})|^2. \tag{4.1}$$

The maximum likelihood estimator is the one that maximizes $L(D, \xi, \boldsymbol{\omega})$. However, $\partial L/\partial \boldsymbol{\omega} = 0$ at $\xi = 0$, so that we cannot analyze the behaviors of the mle by the Taylor expansion of the log likelihood in this case. Following Hartigan (1985) (see also Fukumizu (2000) and Hagiwara et al. (2000) for details), we first fix $\boldsymbol{\omega}$ and search for the ξ that maximizes L. This is easy since L is a quadratic function of ξ. The maximum $\hat{\xi}$ is given by

$$\hat{\xi}(\boldsymbol{\omega}) = \mathrm{argmax}_{\xi} L(D, \xi, \boldsymbol{\omega}) \tag{4.2}$$

$$= \frac{1}{\sqrt{n}} \boldsymbol{a}(\boldsymbol{\omega}) \cdot \tilde{\boldsymbol{x}} = \frac{1}{\sqrt{n}} Y(\boldsymbol{\omega}), \tag{4.3}$$

where

$$\tilde{\boldsymbol{x}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{x}_i. \tag{4.4}$$

By the central limit theorem, $Y(\boldsymbol{\omega}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{a}(\boldsymbol{\omega}) \cdot \boldsymbol{x}_i$ is a Gaussian random variable depending on $\boldsymbol{\omega}$ whose mean is 0 and whose variance is $\boldsymbol{a}(\boldsymbol{\omega}) \cdot \boldsymbol{a}(\boldsymbol{\omega}) = 1$. By substituting $\hat{\xi}(\boldsymbol{w})$ in (4.1), the log likelihood function becomes

$$\hat{L}(\boldsymbol{\omega}) = L(\hat{\xi}(\boldsymbol{\omega}), \boldsymbol{\omega}) \tag{4.5}$$

$$= -\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{x}_i^2 + \frac{1}{2}(\boldsymbol{a}(\boldsymbol{\omega}) \cdot \tilde{\boldsymbol{x}})^2. \tag{4.6}$$

Therefore, the mle $\hat{\boldsymbol{w}}$ is given by the maximizer of $\hat{L}(\boldsymbol{w})$,

$$\hat{\boldsymbol{\omega}} = \mathrm{argmax}_{\boldsymbol{\omega}} \hat{L}(\boldsymbol{\omega}) \tag{4.7}$$

$$= \mathrm{argmax}_{\boldsymbol{\omega}} Y^2(\boldsymbol{\omega}). \tag{4.8}$$

Using the mle, we can obtain the core part of the generalization error of the

form,

$$E_{gen} = E_D E_{\boldsymbol{x}} \left[\log \frac{p_o(\boldsymbol{x})}{p(\boldsymbol{x}|\hat{\xi}, \hat{\omega})} \right] \tag{4.9}$$

$$= E_D E_{\boldsymbol{x}} \left[-\hat{\xi}(\hat{\omega}) \boldsymbol{a}(\hat{\omega}) \cdot \boldsymbol{x} + \frac{1}{2} \hat{\xi}^2(\hat{\omega}) \left(\boldsymbol{a}(\hat{\omega}) \cdot \boldsymbol{a}(\hat{\omega}) \right) \right] \tag{4.10}$$

$$= E_D \left[\frac{1}{2} \hat{\xi}^2(\hat{\omega}) \right] \tag{4.11}$$

$$= \frac{1}{2n} E_D \left[\max_{\omega} Y^2(\omega) \right]. \tag{4.12}$$

Similarly, the training error is obtained by

$$E_{train} = E_D \left[\frac{1}{n} \sum_{i=1}^{n} \log \frac{p_o(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i|\hat{\xi}, \hat{\omega})} \right] \tag{4.13}$$

$$= E_D \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ -\hat{\xi}(\hat{\omega}) \boldsymbol{a}(\hat{\omega}) \cdot \boldsymbol{x_i} + \frac{1}{2} \hat{\xi}^2(\hat{\omega}) \left(\boldsymbol{a}(\hat{\omega}) \cdot \boldsymbol{a}(\hat{\omega}) \right) \right\} \right] \tag{4.14}$$

$$= E_D \left[-\hat{\xi}(\hat{\omega}) \left(\frac{1}{\sqrt{n}} \boldsymbol{a}(\hat{\omega}) \cdot \tilde{\boldsymbol{x}} \right) + \frac{1}{2} \hat{\xi}^2(\hat{\omega}) \right] \tag{4.15}$$

$$= E_D \left[-\frac{1}{2} \hat{\xi}^2(\hat{\omega}) \right] \tag{4.16}$$

$$= -\frac{1}{2n} E_D \left[\max_{\omega} Y^2(\omega) \right]. \tag{4.17}$$

Here, we neglected the common H_o for E_{gen} and E_{train}. One can see the symmetric duality between the generalization error and the training error (Amari and Murata, 1993).

It is in general difficult to calculate the maximum of the Gaussian field $Y(\omega)$. In the simple cone model, we can obtain the explicit value of $E_D \left[\max_{\omega} Y^2(\omega) \right]$. We show the results.

Theorem 1. The generalization and training errors of mle for cone model is given by

$$E_{gen} = \frac{1}{2n(1+c^2)} \left\{ 1 + 2cE\left[|\tilde{x_1}| \right] E\left[|\tilde{\boldsymbol{x}'}| \right] + c^2 E\left[|\tilde{\boldsymbol{x}'}|^2 \right] \right\} \tag{4.18}$$

$$= \frac{1}{2n(1+c^2)} \left\{ 1 + 2c \frac{d!!}{(d-1)!!} \sqrt{\frac{2}{\pi}}^{(-1)^d} + c^2(d+1) \right\}, \tag{4.19}$$

$$E_{train} = -\frac{1}{2n((1+c^2)} \left\{ 1 + 2c\frac{d!!}{(d-1)!!}\sqrt{\frac{2}{\pi}}^{(-1)^d} + c^2(d+1) \right\} \quad (4.20)$$

where $\tilde{\boldsymbol{x}} = (\tilde{x_1}, \tilde{\boldsymbol{x}}') = (\tilde{x}_1, \ldots, \tilde{x}_{d+2})$.

Corollary 1. When d is large, the mle satisfies

$$E_{gen} \approx \frac{c^2 d}{2n(1+c^2)}, \quad (4.21)$$

$$E_{train} \approx -\frac{c^2 d}{2n(1+c^2)}. \quad (4.22)$$

It should be remarked that the generalization and training errors depend on the shape parameter c as well as the dimension number. In the regular case, they depend only on d. As one can easily see, when c is small, the cone looks like a needle, and its behavior resembles a one-dimensional model. When c is large, it resembles two statistical $(d+1)$-dimensional hypersurfaces, so that its behavior is like a d-dimensional regular model.

5. GENERALIZATION ERROR AND TRAINING ERROR OF BAYESIAN PREDICTIVE DISTRIBUTION

While the maximum likelihood estimator searches for an asymptotically optimal point estimator in the model, the Bayes paradigm studies a posterior probability of the parameters based on the set of observations D. The posterior probability density is written as,

$$\begin{aligned} p(\xi, \boldsymbol{\omega}|D) &= c(D)\pi(\xi, \boldsymbol{\omega}) \prod_i p\left(\boldsymbol{x}_i|\xi, \boldsymbol{\omega}\right) \\ &= c(D)\pi(\xi, \boldsymbol{\omega}) \exp\left\{L(D, \xi, \boldsymbol{\omega})\right\}, \end{aligned} \quad (5.1)$$

where $c(D)$ is the normalization factor depending only on data D, $\pi(\xi, \boldsymbol{\omega})$ is a prior distribution on the parameter space, and $\{L(D, \xi, \boldsymbol{\omega})\}$ is the log likelihood of D. The Bayesian predictive distribution $p(\boldsymbol{x}|D)$ is obtained by averaging $p(\boldsymbol{x}|\xi, \boldsymbol{\omega})$ with respect to the posterior distribution $p(\xi, \boldsymbol{\omega}|D)$, and can be written as

$$p(\boldsymbol{x}|D) = \int p(\boldsymbol{x}|\xi, \boldsymbol{\omega}) p(\xi, \boldsymbol{\omega}|D) d\xi d\boldsymbol{\omega}. \quad (5.2)$$

To get an explicit form of the predictive distribution, we need to assume a prior distribution of parameters. One assumes a Gaussian distribution, a uniform

distribution, or the Jeffreys noninformative distribution, some of which might be improper. As long as the prior is a smooth function, the first order asymptotic properties are the same for mle and Bayes estimators in the regular case. However, at singularities, the situation can be different. Here, we assume a uniform prior for $\boldsymbol{\omega}$ that is the regular part of the parameter. For ξ that makes singularity, we assume two different priors, the uniform prior and the Jeffreys prior, and analyzed the two cases, respectively. Using the cone model of section 3.1, we can obtain explicit results for the generalization error and the training error.

Theorem 2. Under the assumption of the uniform prior for ξ, the generalization error and the training error of the predictive distribution is given by

$$E_{gen} = E_D\left[E_{p_o}\left[\log\frac{p_o(\boldsymbol{x})}{p(\boldsymbol{x}|D)}\right]\right] \tag{5.3}$$

$$= \frac{1}{2n}E_D\left[|\boldsymbol{Q}_d^U(\tilde{\boldsymbol{x}})|^2\right], \tag{5.4}$$

where

$$\boldsymbol{Q}_d^U(\tilde{\boldsymbol{x}}) = \nabla S_d^U(\tilde{\boldsymbol{x}}), \tag{5.5}$$

$$S_d^U(\tilde{\boldsymbol{x}}) = \log\int \exp\left\{\frac{(\boldsymbol{a}\cdot\tilde{\boldsymbol{x}})^2}{2}\right\}\pi(\boldsymbol{\omega})d\boldsymbol{\omega}. \tag{5.6}$$

When d is large,

$$E_{gen} = \frac{1}{2n}\left\{1+6\frac{c^4}{d}\right\}. \tag{5.7}$$

$$E_{train} = \frac{1}{n}\sum_{i=1}^{n}E_D\left[\log\frac{p_o(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i|D)}\right] \tag{5.8}$$

$$= E_{gen} - \frac{1}{n}E_D\left[\Delta S_d^U\right], \tag{5.9}$$

and their relationship is given, when d is large, by

$$E_{gen} = \frac{1}{n} + E_{train}. \tag{5.10}$$

The proof needs complicated calculations and is omitted here because of the limitation of the space. We first fix $\boldsymbol{\omega}$ and calculate the integral over ξ, and then calculate the predictive distribution.

The theorem shows rather surprising results : The generalization error decreases as the number d of parameters increases. The relation between E_{gen} and E_{train} does not depend on d, when d is large. These are completely different from the regular case. However, these striking results are given rise to by the uniform prior on ξ. The uniform prior puts strong emphasis on the singularity, showing that one should be very careful for choosing a prior when the model includes singularities.

The Jeffreys prior is uniform in $\boldsymbol{\omega}$ and $\pi(\xi) = |\xi|^d$. This gives a Lebesgue major on the surface of the cone, and looks natural. We show the results in the following theorem whose proof is much more complicated.

Theorem 3. Under the assumption of the Jeffreys prior, the generalization error of the predictive distribution is given by

$$E_{gen} = \frac{1}{2n} E_D \left[|\boldsymbol{Q}_d^J(\tilde{\boldsymbol{x}})|^2 \right]. \tag{5.11}$$

where

$$\boldsymbol{Q}_d^J(\tilde{\boldsymbol{x}}) = \nabla S_d^J(\tilde{\boldsymbol{x}}), \tag{5.12}$$

$$S_d^J(\tilde{\boldsymbol{x}}) = \log \int \pi(\boldsymbol{\omega}) I_d(\boldsymbol{a} \cdot \tilde{\boldsymbol{x}}) \exp\left\{ \frac{(\boldsymbol{a} \cdot \tilde{\boldsymbol{x}})^2}{2} \right\} d\boldsymbol{\omega}, \tag{5.13}$$

$$I_d(u) = \frac{1}{\sqrt{2\pi}} \int |z+u|^d e^{(-\frac{z^2}{2})} dz. \tag{5.14}$$

$$\tag{5.15}$$

When d is large, the generalization error increases in proportion to d. We can calculate the training error similarily.

6. CONCLUSIONS AND DISCUSSIONS

We have analyzed the asymptotic behaviors of the MLE and Bayes estimators in terms of the generalization error and the training error by using a simple statistical model (cone model), when the true parameter is unidentifiable. Since the classic paradigm of statistical inference based on the Cramér-Rao theorem does not hold in such a singular case, we need a new theory. By analyzing the relationship between the generalization error and the training error, we can obtain a basic criterion for model selection. We can also compare the estimation accuracy of the maximum likelihood estimator and the Bayesian predictive distri-

bution from the results of analysis. Under the proposed framework, the various estimation methods can be studied and compared to each other.

REFERENCES

Amari, S. (1998). Natural gradient works efficiently in learning, *Neural Computation,* **10**, 251-276.

Amari, S. and Murata, N. (1993). Statistical theory of learning curves under entropic loss criterion *Neural Computation,* **5**, 140-153.

Amari S. and Nagaoka, H. (2000). *Methods of Information Geometry,* AMS and Oxford University Press.

Amari, S. and Ozeki, T. (2001). Differential and algebraic geometry of multilayer perceptrons, *IEICE Transactions on Fundamentals of Electronics, Communications and Computer System,* **E84-A**, 31-38.

Amari, S., Park, H., and Fukumizu, F. (2000). Adaptive method of realizing natural gradient learning for multilayer perceptrons, *Neural Computation,* **12**, 1399-1409.

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

Fukumizu, K. (2000). Statistical analysis of unidentifiable models and its application to multilayer neural networks, *Memo at Post-Conference of the Bernoulli-RIKEN BSI 2000 Symposium on Neural Networks and Learning.*

Fukumizu, K. (2001). Likelihood Ratio of Unidentifiable Models and Multilayer Neural Networks, *Research Memorandum,* **780**, Inst. of Statistical Mathematics.

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.*

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures, *Proceedings of Berkeley Conference in Honor of J. Neyman and J. Kiefer,* **2**, 807-810.

Park, H., Amari, S. and Fukumizu, F. (2000). Adaptive natural gradient learning algorithms for various stochastic models, *Neural Networks*, **13**, 755-764.

Watanabe, S. (2001a). Algebraic analysis for non-identifiable learning machines, *Neural Computation*, **13**, 899-933.

Watanabe, S. (2001b). Training and generalization errors of learning machines with algebraic singularities (in Japanese), *The Trans. of IEICE A*, **J84-A**, 99-108.