(19)「統計推測の理論とその応用—幾何学的方法と特異モデル—」に関する研究報告

Generalization Error and Training Error in Some Simple Singular Statistical Models

Shun-ichi Amari, Hyeyoung Park, Tomoko Ozeki
RIKEN Brain Science Institute

1 Introduction

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. From the geometrical point of view, distributions specified by unidentifiable parameters become singular points 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[1, 2, 3, 4, 5].

In the present report, we first demonstrate a method of analyzing the generalization and training errors for mle and Bayes predictive distribution in terms of Gaussian random fields, by using a simple cone models. The obtained results are completely different from regular statistical models where Carmér-Rao paradigm holds, and the method can be applied to the multilayer perceptron. We then show another approach of analyzing the asymptotic performance at singularities by using a simple Gaussian mixture model.

2 Singular Statistical Models

In the present report, we use simple toy models to analyze the singularities. Let us first introduce a simple cone model: Let $\boldsymbol{x}$ be Gaussian random variable $\boldsymbol{x} \in R^{d+2}$, with mean $\boldsymbol{\mu}$ and identity covariance matrix $\boldsymbol{I}$, and let $S = \{\boldsymbol{\mu} | \boldsymbol{\mu} \in R^{d+2}\}$ be the parameter space. The cone model M is a subset of S, embedded as

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

where c is a constant, $||\boldsymbol{a}^2|| = 1$, $\boldsymbol{\omega} \in S^d$ and S^d is a d-dimensional unit sphere. The M is a cone, having $(\xi, \boldsymbol{\omega})$ as coordinates, where the apex $\xi = 0$ is the singular point.

A simple multilayer perceptron has also the same singular structure. The input-output relation of a simple multilayer perceptron is given by

$$y = v\varphi(\boldsymbol{w} \cdot \boldsymbol{x}) + n \tag{2}$$

When $v = 0$, the behavior is the same whatever $\boldsymbol{w}$ is, which makes singularity.

We also discuss about a simple Gaussian Mixture model of the form,

$$p(x; v, w_1, w_2) = (1-v)\phi(x - w_1) + v\phi(x - w_2), \tag{3}$$

where $\phi(x)$ is the standard Gaussian density function. The singularities occur at $v(1-v)(w_1 - w_2) = 0$.

3 Analysis of Cone Model and MLP

For the cone model, we define the Gaussian random field, $Y(\boldsymbol{\omega}) = \boldsymbol{a}(\boldsymbol{\omega}) \cdot \tilde{\boldsymbol{x}}$, where $\tilde{\boldsymbol{x}} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \boldsymbol{x}_t$. Then Following Hartigan [5] (see also [3] for details), we can obtain following results for the case of mle.

Theorem 1. the generalization and training error of mle is given by

$$E_{gen} = E_D E_x \left[\log \frac{p_o(x)}{p(x|\hat{\xi}, \hat{\omega})} \right] = \frac{1}{2T} E_D \left[\sup_{\omega} Y^2(\omega) \right] \approx \frac{c^2 d}{2T(1+c^2)}, \tag{4}$$

$$E_{train} = E_D \left[\frac{1}{n} \sum_{i=1}^{n} \log \frac{p_o(x_i)}{p(x_i|\hat{\xi}, \hat{\omega})} \right] = -\frac{1}{2T} E_D \left[\sup_{\omega} Y^2(\omega) \right] \approx -\frac{c^2 d}{2T(1+c^2)}. \tag{5}$$

For the Bayes prediction, we can also use the Gaussian random filed $Y(\omega)$, and get the following results.

Theorem 2. Under the Jeffreys prior for ξ, the generalization error and the training error of the predictive distribution are given by

$$E_{gen} = \frac{1}{2T} E_D \left[\|\nabla \log P_d(\tilde{x})\|^2 \right], \quad E_{train} = E_{gen} - \frac{1}{T} E_D \left[\nabla \log P_d(\tilde{x}) \cdot \tilde{x} \right], \tag{6}$$

$$I_d(u) = \frac{1}{\sqrt{2\pi}} \int |z+u|^d \exp\left\{ -\frac{1}{2} z^2 \right\} dz, P_d(\tilde{x}) = \int I_d(Y(\omega)) \exp\left\{ \frac{1}{2} Y^2(\omega) \right\} d\omega. \tag{7}$$

Under the uniform prior, the above results hold by replacing $I_d(Y)$ by 1. In addition, we obtained $E_{gen} = (d+1)/2T$ for the Jeffreys prior, and $E_{gen} = 1/2T$ for the uniform prior.

For the simple MLP model defined in (2), we can also apply the same Gaussian random field approach, and get similar results. (See [6] for details.)

4 Analysis of Gaussian Mixture

For the Gaussian mixture model defined in (3), we consider the case were $w_2 \geq w_1$, and $u = w_2 - w_1 \geq 0$ is small. We treat the case where v or $1-v$ is not very small. We introduce the new coordinates, $u = w_2 - w_1$, $w = (1-v)w_1 + vw_2$, where u indicating the difference of the two peaks and w their center of mass.

For fixed w, let us introduce an exponential family $S_w = \{p(z; \alpha, \beta)\}$ of the form,

$$p(z; \alpha, \beta) = \exp\left\{ -\frac{z^2}{2} + \frac{\alpha}{\sqrt{2}}(z^2 - 1) + \frac{\beta}{\sqrt{6}}(z^3 - 3z) - \psi(\alpha, \beta) \right\}, \tag{8}$$

under a suitable measure, where we put $z = x - w$.

When u is small, by neglecting higher order terms, our model is embedded in the regular model S_w by

$$\alpha = \frac{v(1-v)}{\sqrt{2}} u^2, \quad \beta = -\frac{1}{\sqrt{6}} v(1-v)(2v-1)u^3, \tag{9}$$

which is singular. This shows the cusp type algebraic singularity. Then we can obtain that, when u is small and the number of observations T is large,

$$\Delta u \sim \frac{1}{\sqrt{T} u^2}, \quad \Delta v \sim \frac{1}{\sqrt{T} u^3}. \tag{10}$$

References

[1] Hagiwara, K., Hayasaka, K., Toda, N., Usui, S., and Kuno, K. (2001). *Neural Networks*, **14** 1419-1430.

[2] Watanabe, S. (2001). *Neural Computation*, **13**, 899-933.

[3] Fukumizu, K. (2001). *Research Memorandum*, **780**, Inst. of Statistical Mathematics.

[4] Dacunha-Castelle, D. and Gassiat, E. (1997). *Probability and Statistics*, **1**, 285-317.

[5] Hartigan, J. A. (1985). *Proceedings of Berkeley Conference in Honor of J. Neyman and J. Kiefer*, **2**, 807-810.

[6] Amari, S., Park, H., and Ozeki, T. (2002) *Proceedings of NIPS2001*, to Appear.

Componentwise shrinkage in orthogonal regression

三重大学 工学部 萩原克幸

1 Introduction

In harmonic analysis, we often encounter the problem of determining mainly contributed frequency components. The problem is formulated as model selection of an orthogonal regression. From a stand point of the harmonic analysis, it is natural to estimate the frequency components whose contributions are large. Then, we should consider the orthogonal regression model, in which we estimate not only the coefficients but also the components in the parameter estimation. Therefore, the orthogonal regression model is not linear one and becomes nonlinear. Indeed, the model is found to be without identifiability. [3] and [4] have proposed a model selection criterion for the problem independently and [3] has shown that the criterion has a consistency of model selection.

On the other hand, the minimization of the cost function which consists of the empirical error plus a penalty term is known to yield the shrinkage estimator and it is possible to improve prediction performance compared with the least squares estimator. In this paper, we consider the introduction of the componentwise shrinkage into the orthogonal regression mentioned in the above. Then, we extend the Quinn–Sakai's criterion to handle the componentwise shrinkage estimator.

2 Formulation of the problem

Let us denote N pairs of input–output data by $\left\{(\boldsymbol{x}_n, y_n) : \boldsymbol{x}_n \in \mathbf{R}^d, y_n \in \mathbf{R}, 1 \leq n \leq N\right\}$. Here, output data y_n is generated according to $y_n = h(\boldsymbol{x}_n) + \xi_n$, where h is a true function and ξ_n is additive noise. Throughout this paper, we assume that $\xi_1, \ldots, \xi_n$ are independent samples according to a Gaussian distribution $N(0, \sigma^2)$. Here, we consider the regression with $f_{\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x}_n) = \sum_{k=1}^{K} a_k g_{b_k}(\boldsymbol{x}_n)$, where $\boldsymbol{a} = (a_1, \ldots, a_K), a_k \in \mathbf{R}$ is the coefficient vector and $\boldsymbol{b} = (b_1, \ldots, b_K), b_k \in \{1, \ldots, N\}$ is the index parameters which determine the components. Unlike the linear orthogonal regression, this function has a parameter $\boldsymbol{b}$. To estimate $\boldsymbol{b}$ corresponds to the choice of components from the family $\mathcal{G} = \{g_1, \ldots, g_N\}$ in the estimation procedure. Here, we assume that the orthogonality condition for the elements in $\mathcal{G}$ as $\sum_{n=1}^{N} g_k(\boldsymbol{x}_n) g_l(\boldsymbol{x}_n)$ is equal to s_k if $k = l$ and 0 if $k \neq l$. Now, the cost function is defined by $C(\boldsymbol{a}, \boldsymbol{b}, \lambda) = r_{\mathrm{emp}}(\boldsymbol{a}, \boldsymbol{b}) + \sum_{k=1}^{K} \lambda_k a_k^2$, where $r_{\mathrm{emp}}(\boldsymbol{a}, \boldsymbol{b}) = \sum_{n=1}^{N} \left(y_n - f_{\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x}_n)\right)^2 / N$ is the empirical squared error for $f_{\boldsymbol{a},\boldsymbol{b}}$. Let us define $Z_{b_k} = (\lambda_{b_k} + s_{b_k}) \overline{a}_{b_k}^2$ and $\overline{a}_{b_k} = \sum_n y_n g_{b_k}(\boldsymbol{x}_n) / (\lambda_{b_k} + s_{b_k})$. Let $l_1, \ldots, l_N$ be the indexes which satisfy $Z_{l_1} \geq Z_{l_2} \geq \cdots \geq Z_{l_N}$. Then, we have $\widehat{b}_k = l_k$ and $\widehat{a}_k = \overline{a}_{l_k}$ as the minimizing parameters of the cost. Let us define $\rho_{b_k} = s_{b_k} / (\lambda_{b_k} + s_{b_k})$. Then, the least squares estimator of a_{b_k} is given by $\widetilde{a}_{b_k} = \sum_n y_n g_{b_k}(\boldsymbol{x}_n) / s_{b_k}$ and we have $\overline{a}_{b_k} = \rho_{b_k} \widetilde{a}_{b_k}$. Because $0 \leq \rho_{b_k} \leq 1$, $\overline{a}_{b_k}$ is a shrinkage estimator, which is assigned to each component independently.

3 Overfitting property for noise

In the following, to see the overfitting property of $f_{\boldsymbol{a},\boldsymbol{b}}$ for noise, we assume that $y_n = \xi_n$, $n = 1, \ldots, N$; i.e. the data is a Gaussian noise sequence. Let $z_1, \ldots, z_N$ be i.i.d. samples from the distribution of y_1 and $y_1, \ldots, y_N, z_1, \ldots, z_N$ be independent. Let us define the prediction error by $r(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{b}}) = \sum_{n=1}^{N} \mathbf{E}_{\boldsymbol{z}} \left(z_n - f_{\widehat{\boldsymbol{a}},\widehat{\boldsymbol{b}}}(\boldsymbol{x}_n)\right)^2 / N$, where $\boldsymbol{z} = (z_1, \ldots, z_N)$ and $\mathbf{E}_{\boldsymbol{z}}$ denotes the expectation with respect to the joint distribution of $\boldsymbol{z}$. Then we have the following theorems.

Theorem 1 *Let us define* $\overline{C}_N = 2 \log N + (-1 + \epsilon) \log \log N$, *where* ϵ *is an arbitrary positive constant. Then, for any fixed* K, *we have* $\lim_{N \to \infty} \mathbf{P}\left\{r(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{b}}) \leq r_{\mathrm{emp}}(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{b}}) + \frac{2}{N} \sum_{k=1}^{K} \rho_{l_k} \overline{C}_N + \delta\right\} = 1$ *for any* $\delta > 0$.

Theorem 2 *Let us define* $\underline{C}_N = 2 \log N + (-1 - \epsilon) \log \log N$ *and assume that* $\rho_l > \gamma > 0$ *for any* l. *Then, for any fixed* K, $\lim_{N \to \infty} \mathbf{P}\left\{r(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{b}}) > r_{\mathrm{emp}}(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{b}}) + \frac{2K\gamma}{N} \underline{C}_N - \delta\right\} = 1$ *for any* $\delta > 0$.

4 Model selection criterion

Theorem 1 tells us that the correction on the overfitting to noise is not more than $2\sum_{k=1}^{K}\rho_{l_k}\overline{C}_N$ and this is tight because of Theorem 2 for large N. Then, we proposed the model selection criterion whose form is given by $\mathrm{MSC}(K) = r_{\mathrm{emp}}(\widehat{a}_K, \widehat{b}_K) + \frac{2\sigma^2}{N}\sum_{k=1}^{K}\rho_{l_k}\overline{C}_N$. Here, if $\rho_l = 1$ for $l = 1, \ldots, N$ then the criterion reduces to Hayasaka's criterion[2], which is given by $\mathrm{MSC}_0(K) = r_{\mathrm{emp}}(\widetilde{a}_K, \widetilde{b}_K) + \frac{2\sigma^2 K}{N}C_N$, where $r_{\mathrm{emp}}(\widetilde{a}_K, \widetilde{b}_K) = \min_{a,b} r_{\mathrm{emp}}(a, b)$. Note that $\mathrm{MSC}_0(K)$ is equivalent with Quinn–Sakai's criterion[4][3]. It has been shown that MSC_0 has the consistency of model selection. Also we can show the consistency of MSC if we assume that $0 < \rho_l \leq 1$ and $s_l = O(N)$ for all l. This is because $\mathrm{MSC}(K) = O_p(1)$ for $K < K^*$ and $\mathrm{MSC}(K^*) = o_p(1)$, where K^* is the true number of components, and $\mathrm{MSC}(K) > \mathrm{MSC}(K^*)$ for $K > K^*$ due to Theorem 1. On the other hand, in practical situations, we should estimate ρ_l. The minimizing ρ_l of the expected prediction error is given by $\rho_l^* = \alpha_l^2 s_l/(\sigma^2 + \alpha_l^2 s_l)$, where α_l^2 is the true coefficient of the lth component. Then, we consider the estimate $\widehat{\rho}_k = s_k\widetilde{a}_k^2/(\widehat{\sigma}^2 + s_k\widetilde{a}_k^2)$, where $\widehat{\sigma}^2$ is an appropriate estimator of $\widehat{\sigma}^2$. $\widehat{\rho}_k$ has also been employed in [1] as an empirical estimate of the shrinkage parameter in designing the digital Wiener filter.

5 Numerical simulation

In the simulation to evaluate the performance of MSC, $\mathcal{G}$ is set to be an orthogonal family whose components consist of sinusoidal functions with different frequency components. Here, we set $h(x) = \sum_{k=1}^{10} g_k(x)$ and $\sigma^2 = 1$. In the simulation, we estimate the parameter for each of 500 sets of data with size N. For each estimation, the prediction error is estimated by the average squared error on 1000 sets of new data with size N. The averaged prediction error is calculated as the average of the estimated prediction errors for 500 trials. As the number of data, we take $N = 50, 100, 200, 400$. Table 1 shows the averaged prediction error at the selected number of components by MSC_0 and MSC. As we can see, MSC outperform MSC_0 at any N.

6 Conclusions and Future works

In this paper, we consider a model selection criterion for an orthogonal regression with componentwise shrinkage estimators. By a numerical simulation, we showed that the proposed criterion with the empirical estimate of the amount of the shrinkage exhibits better performance compared with the criterion under the least squares estimation. As seen in this paper, the proposed criterion, also the criterion proposed in [3][4][2], are constructed based only on the overfitting property for noise. Indeed, for our regression model, it may be shown that the variance of the estimated coefficient of the true component is $O(1)$ but that of the other component is $O(\log N)$ for large N. Therefore, the penalty terms of the criteria may be overestimation. This problem emerges in considering model selection of any regression models without identifiability. One possible way to overcome this problem is to suppress the variance only for the overfitting to noise by using the componentwise shrinkage and construct a criterion under this setting. This is left as a future work.

N	50	100	200	400
MSC_0	1.41947	1.29090	1.21088	1.14102
MSC	1.40364	1.28068	1.20411	1.13537

Table 1: The averaged prediction errors of the models, which is selected by MSC_0 and MSC.

References

[1] S. Ghael, A. M. Sayeed and R.G. Baraniuk, in *Proceedings of SPIE*, San Diego, 389–399, 1997.

[2] T. Hayasaka, K. Hagiwara, N. Toda and S. Usui, in *Proceedings of 1999 Workshop on Information-Based Induction Sciences, IBIS'99*, Izu, Japan, 151–156, 1999.

[3] B. G. Quinn, *J. Time Ser. Anal.*, **10**, 71–75, 1989.

[4] H. Sakai, *IEEE Trans. Acoust., Speech, Signal Processing*, **38**, 999–1004, 1990.

Gaussian volumes of tubes and Euler characteristic densities

Stanford Univ. Jonathan Taylor

In this work we describe some new results in approximating the distribution of the maximum of a smooth stochastic process on a manifold M, specifically Gaussian and closely related processes. We use the expected Euler characteristic method to approximate this distribution and describe a new formula, a Gaussian version of the classical Kinematic Fundamental Formulae of integral geometry. This result relates the Euler characteristic densities of a certain class of processes to coefficients in certain power series expansions of the standard Gaussian measure of certain tubes.

We give some simple applications of the result, including a simple derivation of the EC densities of Gaussian and χ^2 processes. Although these two results are not new, the Gaussian KFF sheds some light on them gives a geometric interpretation of them. As a corollary to the result for χ^2 processes, the Gaussian KFF yields the EC densities of non-central χ^2 processes, which have not been published previously.

This Gaussian KFF also shows how to use the EC approach for certain fields with piecewise smooth level sets. As an application, we derive the EC densities of the process given by taking the pointwise minimum of two i.i.d. Gaussian processes, known as a correlated conjunction. To validate these approximations, we conclude with the results of a small simultation study.

Tail probability via tube formula when critical radius is zero

竹村 彰通 (A.Takemura) 東京大学大学院情報理工学系研究科
栗木 哲 (S.Kuriki) 統計数理研究所

Let M be a closed subset of the unit sphere S^{n-1} in R^n. We consider upper tail probability of the maximum of a random field $Z(u)$, $u = (u_1, \ldots, u_n)' \in M$, defined by

$$Z(u) = u'z = \sum_{i=1}^{n} u_i z_i,$$

where $z = (z_1, \ldots, z_n)'$ is distributed according to n-dimensional standard multivariate normal distribution $N_n(0, I_n)$. This is the canonical form of Gaussian random field with finite Karhunen-Loève expansion and constant variance as discussed in Takemura and Kuriki (2002). Let $y = (y_1, \ldots, y_n)' = z/\|z\|$ be distributed according to the uniform distribution $\mathrm{Unif}(S^{n-1})$ on the unit sphere S^{n-1}. We also study upper tail probability of the maximum of

$$Y(u) = u'y.$$

In Takemura and Kuriki (1997) we treated convex M for studying the properties of $\bar{\chi}^2$ distribution in the framework of testing against multivariate ordered alternatives. In Kuriki and Takemura (2001) we treated smooth M without boundary for studying multilinear forms in normal variates. Unifying these cases in Takemura and Kuriki (2002) we considered index set M which is locally approximated by a convex cone. We established that in this case M has positive critical radius and the tube method by Sun (1993) and the Euler characteristic method by Adler (1981) and Worsley (1995a, b) lead to identical valid asymptotic expansion of the upper tail probabilities. In a different setting, Adler (2000) showed that the Euler characteristic method for isotropic Gaussian random fields on piecewise smooth domain gives valid asymptotic expansion using the results by Piterbarg (1996).

These results might give an impression that the formal asymptotic expansion based on the tube formula is valid and identical to the Euler characteristic method for practically all regular cases. However this is not the case if the critical radius of M is zero. The main purpose of this paper is to show that if the critical radius of M is zero, the asymptotic expansion based on the tube formula is generally incorrect except for the main term of the expansion. Furthermore the equivalence of the formal tube formula and the Euler characteristic method no longer holds. We also give some simple examples of index sets with zero critical radius, for which the formal tube formula and the Euler characteristic method give different asymptotic expansions and both are incorrect. More substantial application of the results of the present paper is given in Takemura and Kuriki (2001), where a natural multivariate test statistics has an associated index set with zero critical radius.

References

[1] Adler, R. J. (1981). *The Geometry of Random Fields*. Wiley, Chichester.

[2] Adler, R. J. (2000). On excursion sets, tube formulas and maxima of random fields. *Ann. Appl. Probab.*, **10**, 1–74.

[3] Cao, J. and Worsley, K. J. (1998). The geometry of correlation fields with an application to functional connectivity of the brain. *Ann. Appl. Probab.*, **9**, 1021–1057.

[4] Cao, J. and Worsley, K. J. (1999). The detection of local shape changes via the geometry of Hotelling's T^2 fields. *Ann. Statist.*, **27**, 925–942.

[5] Cheeger, J., Müller, W. and Schrader, R. (1986). Kinematic and tube formulas for piecewise linear spaces. *Indiana University Mathematics Journal*, **35**, 737–754.

[6] Kuriki, S. and Takemura, A. (2001). Tail probabilities of the maxima of multilinear forms and their applications. *Ann. Statist.*, **29**, 328–371.

[7] Matheron, G. (1975). *Random Sets and Integral Geometry*. Wiley, New York.

[8] Milnor, J. (1963). *Morse Theory*. Ann. Math. Studies, **51**, Princeton University Press, Princeton.

[9] Piterbarg, V. I. (1996). *Asymptotic Methods in the Theory of Gaussian Processes and Fields*. Translations of Mathematical Monographs, **148**, American Mathematical Society, Providence.

[10] Santaló, L. A. (1976). *Integral Geometry and Geometric Probability*. Addison-Wesley, London.

[11] Schneider, R. (1993). *Convex Bodies: The Brunn-Minkowski Theory*. Cambridge University Press, Cambridge.

[12] Stoyan, D., Kendall, W. S. and Mecke, J. (1995). *Stochastic Geometry and Its Applications*. 2nd ed. Wiley, New York.

[13] Sun, J. (1993). Tail probabilities of the maxima of Gaussian random fields. *Ann. Probab.*, **21**, 34–71.

[14] Takemura, A. and Kuriki, S. (1997). Weights of $\bar{\chi}^2$ distribution for smooth or piecewise smooth cone alternatives. *Ann. Statist.*, **25**, 2368–2387.

[15] Takemura, A. and Kuriki, S. (2001). Maximum covariance difference test for equality of two covariance matrices. In *Algebraic Methods in Statistics and Probability*, pp. 283–302, M. Viana and D. Richards eds., Contemporary Mathematics, **287**, American Mathematical Society, Providence.

[16] Takemura, A. and Kuriki, S. (2002). On the equivalence of the tube and Euler characteristic methods for the distribution of the maximum of Gaussian fields over piecewise smooth domains. *Ann. Appl. Probab.*, to appear.

[17] Worsley, K. J. (1994). Local maxima and the expected Euler characteristic of excursion sets of χ^2, F and t fields. *Adv. Appl. Probab.*, **26**, 13–42.

[18] Worsley, K. J. (1995a). Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets, with applications to medical images. *Ann. Statist.*, **23**, 640–669.

[19] Worsley, K. J. (1995b). Boundary corrections for the expected Euler characteristic of excursion sets of random fields, with an application to astrophysics. *Adv. Appl. Probab.*, **27**, 943–959.

Tube method and Euler characteristic method for Gaussian random fields with inhomogeneous variance

Satoshi Kuriki (Inst. Statist. Math.)
Akimichi Takemura (Univ. Tokyo)

1. Tube with nonconstant radius.

Define a Gaussian field with finite Karhunen-Loève expansion and its standardized version:

$$X(t) = \phi(t)'z, \quad Y(t) = \phi(t)'z/\|z\|, \quad t \in I \subset R^d,$$

where $z \sim N_p(0, I_p)$ is a standard p-dimensional Gaussian random vector, and $\phi : I \to \phi(I) \subset R^p$ is a one-to-one C^2-mapping. We assume that the image $\phi(I)$ is a compact C^2-submanifold of R^p without boundary not containing the origin. Let $\sigma(t) = \|\phi(t)\|$ and $\varphi(t) = \phi(t)/\sigma(t)$. We also assume that $\varphi(t) : I \to S^{p-1}$ a one-to-one C^2-mapping, where S^{p-1} denotes the unit sphere in R^p.

$X(t)$ is a continuous Gaussian filed with mean zero, variance $\sigma^2(t)$, and the correlation function $\mathrm{corr}(t_1, t_2) = \varphi(t_1)'\varphi(t_2)$, t_1, $t_2 \in I$.

Consider the maxima over the index set I,

$$T = \max_{t\in I} X(t) = \max_{t\in I} \phi(t)'z, \qquad U = \max_{t\in I} Y(t) = \max_{t\in I} \phi(t)'z/\|z\|.$$

They exist because of the assumption of the compactness of $\phi(I)$. In this paper we will discuss upper tail probabilities of the distributions of T and U.

Put

$$M = \{\varphi(t) \mid t \in I\}. \tag{1}$$

Since $\|\varphi(t)\| = 1$, M is a C^2-submanifold of the unit sphere $S^{p-1} \subset R^p$. The function σ can be regarded as a function on M by $\sigma(u) := \sigma(\varphi^{-1}(u))$, $u \in M$. The distributions of T and U depend only on the manifold M and the function $\sigma : M \to R$.

Noting that the independence of $\|z\|$ and $z/\|z\|$, we have

$$P(T > x) = E\left[P\left(U > \frac{x}{\|z\|} \mid \|z\|\right)\right]. \tag{2}$$

Here

$$\begin{aligned} P(U > a) &= P\left(\max_t \sigma(t)\varphi(t)'z/\|z\| > a\right) = P\left(\exists t,\ \varphi(t)'z/\|z\| > a/\sigma(t)\right) \\ &= P\left(\exists t,\ \cos^{-1}(\varphi(t)'z/\|z\|) < \cos^{-1}(a/\sigma(t))\right). \end{aligned}$$

Since $z/\|z\|$ is distributed uniformly on S^{p-1}, $P(U > a)$ is $1/\mathrm{Vol}(S^{p-1})$ times the volume of tube around M with generally nonconstant radius:

$$\left\{u \in S^{p-1} \mid \exists t,\ \cos^{-1}(\varphi(t)'u) < \cos^{-1}(a/\sigma(t))\right\}.$$

2. Main Results.

Here we use conventions $\partial/\partial t^i = \partial_i$, $\varphi_i = \partial_i\varphi$, $\varphi_{ij} = \partial_i\partial_j\varphi$, *etc.* The manifold M in (1) is a Riemannian manifold with the metric $g_{ij} = \varphi_i'\varphi_j$, and the affine connection $\Gamma_{ij,k} = \varphi_{ij}'\varphi_k$. The covariant derivative is denoted by ∇. Let $c_{ij} = -\nabla_i\nabla_j\ell + \nabla_i\ell\nabla_j\ell$, $\ell = \log\sigma$, be a $(0,2)$ symmetric tensor. Let R_{ijkl} be the curvature tensor, and let $\tilde{R}_{ijkl} = R_{ijkl} - (g_{ik}g_{jl} - g_{il}g_{jk})$. Denote (i,j)-th element of the inverse matrix of $d_{ij} = g_{ij} + c_{ij}$ by d^{ij}. Then we have the following:

Theorem 1 *(volume of tube).*

$$\begin{aligned} P(U > x) &= \int_I \det(g_{ij} + c_{ij}) \det(g_{ij})^{-\frac{1}{2}} \bigwedge_i dt^i \\ &\quad \times \sum_{e=0}^{d} \frac{\Gamma(\frac{d-e+1}{2})}{2^{1+\frac{e}{2}}\pi^{\frac{1}{2}(d+1)}} (1 + \|\nabla\ell\|^2)^{-\frac{1}{2}(d-e+1)} \\ &\quad \times \bar{B}_{\frac{1}{2}(d-e+1),\frac{1}{2}(p-d+e-1)}\left(\frac{1+\|\nabla\ell\|^2}{\sigma^2}x^2\right) \times \zeta_e(t) \end{aligned} \tag{3}$$

for $x \geq \max_{u\in M}\sigma(u)\cos\theta_c$, where θ_c is a positive constant. $\bar{B}_{a,b}(\cdot)$ is upper probability of the beta distribution with parameters (a,b). $\zeta_e(t) = 0$ for e odd, and for e even

$$\zeta_e(t) = \sum_{1\leq i_1<\cdots<i_e\leq p} \sum_{[j,k]} \varepsilon[j,k]\, \tilde{R}_{j_1j_2k_1k_2}\cdots\tilde{R}_{j_{e-1}j_ek_{e-1}k_e}\, d^{j_1k_1}\cdots d^{j_ek_e},$$

where the summation $\sum_{[j,k]}$ is taken over all possible pairings $\{(j_1,j_2),\ldots,(j_{e-1},j_e)\}$ and $\{(k_1,k_2),\ldots,(k_{e-1},k_e)\}$ from $\{i_1,\ldots,i_e\}$ such that $j_1 < j_2$, ..., $j_{e-1} < j_e$, $k_1 < k_2$, ..., $k_{e-1} < k_e$, and $j_1 < j_3 < \cdots < j_{e-1}$. $\varepsilon[j,k] = \mathrm{sgn}(j_1,\ldots,j_e;k_1,\ldots,k_e)$ is the sign of permutation. In particular, $\zeta_0(t) = 1$ and $\zeta_2(t) = \frac{1}{2}\tilde{R}_{ijkl}d^{ik}d^{jl}$.

Once the distribution of U is obtained, the distribution of T can be derived via (2), that is, by substituting $x^2 := x^2/\|z\|^2$ in (3), and taking expectation with respect to $\|z\|^2 \sim \chi^2(p)$. As in Sun (1993) and Kuriki and Takemura (2001), we can prove the following theorem.

Theorem 2 *(tube method).* $P(T > x) = \tilde{P}(T > x) + O(\bar{G}_p(x^2(1+\tan^2\theta_c)))$, $x \to \infty$, *where*

$$\begin{aligned} \tilde{P}(T > x) &= \int_I \det(g_{ij} + c_{ij}) \det(g_{ij})^{-\frac{1}{2}} \bigwedge_i dt^i \\ &\quad \times \sum_{e=0}^{d} \frac{\Gamma(\frac{d-e+1}{2})}{2^{1+\frac{e}{2}}\pi^{\frac{1}{2}(d+1)}} (1 + \|\nabla\ell\|^2)^{-\frac{d-e+1}{2}} \\ &\quad \times \bar{G}_{d-e+1}\left(\frac{1+\|\nabla\ell\|^2}{\sigma^2}x^2\right) \times \zeta_e(t), \end{aligned}$$

and $\bar{G}_\nu(\cdot)$ is upper probability of the χ^2 distribution with ν degrees of freedom.

Theorem 3 *(Laplace approximation). Assume that $\sigma(t)$ takes the unique maximum at $t = 0$. Without loss of generality assume $\sigma(0) = 1$. Then*

$$P(T > x) \sim \bar{\Phi}(x) \times \det(g_{ij}(0) - \ell_{ij}(0))^{\frac{1}{2}} \det(-\ell_{ij}(0))^{-\frac{1}{2}}, \quad x \to \infty,$$

where $\bar{\Phi}(u)$ is upper probability of the standard normal distribution $N(0,1)$.

Theorem 4 *(Equivalence of the tube method and the EC method).* Omitted.

The Multivariate Analog of the One-Sided Test Revisited

Akio KUDO (Hyogo University)
and Yoshiro YAMAMOTO (Tama University)

1 INTRODUCTION

It is already nearly four decades since the paper of Kudo([2]) was published 1963. The problem was as follows. Given a p-variate normal distribution: $N_p(\boldsymbol{\theta}, \sigma^2\Lambda)$, we consider the testing problem: $\mathrm{H}_0 : \boldsymbol{\theta} = \mathbf{0}$ versus $\mathrm{H}_1 : \boldsymbol{\theta} \geq 0$, where Λ is assumed to be known, but the scale parameter σ^2 is unknown. $\boldsymbol{\theta}' = (\theta_1, \cdots, \theta_p) \geq \mathbf{0}$ means Max $\theta_i > 0$, Min $\theta_i \geq 0$. In Kudo's original paper [2], the unknown scale parameter was absent, but it was immediately added in the formulation in the book by Barlow, Bartholomew, Bremner and Brunk [1].

In this book, published in 1972, it is stated that the algorithm for computating the test statistic is not available, and mreover the null distribution "may be expressed in terms of orthant probabilities of certain multivariate normal distribution, and no general solution in closed form exists for $r > 3$. Thus the range of the solutions in which of test can be used is rather severely restricted." (see on page 177-8 in [1] In short they pointed our two difficulties; one lies in computation of the statistic and other is how to compute the null distribution.

After about 15 years later another book [6] was published. This book refers to the paper [2] more frequently but the authors of this book were not seem to pay attention to the difficulties stated in [1].

The purpose of this presentation is to describe the developments in the area related to the above dificulties.

They assumed there exists an independent estimator T^2 of σ^2 based on χ^2 distribution. The main purpose of this paper is to present a not widely recognized fact that the existence of T^2 is not necessary needed.

2 THE LIKELIHOOD RATIO TEST

The maximum likelihood estimate (MLE) can be obtained by the method reviewed in [9], and let the MLE be $\boldsymbol{X}^0$. We have the partition of the quadratic form; $\boldsymbol{X}'\Lambda^{-1}\boldsymbol{X} = \boldsymbol{X}^{0'}\Lambda^{-1}\boldsymbol{X}^0 + S^0$. and the likelihood ratio test rejects H_0 when $\dfrac{\boldsymbol{X}^{0'}\Lambda^{-1}\boldsymbol{X}^0}{S^0 + T^2}$ is too large. This statistic may take the value 0, as $\boldsymbol{X}^0 = \mathbf{0}$ may hold true, and moreover the value 1 when T^2 is not present and $\boldsymbol{X}^0 = \boldsymbol{X}$.

3 DISTRIBUTION

We discuss the distribution of the following statistic.

$$\bar{E}_n^2 = \frac{\boldsymbol{X}^{0'}\Lambda^{-1}\boldsymbol{X}^0}{\boldsymbol{X}'\Lambda^{-1}\boldsymbol{X} + T^2} = \frac{\boldsymbol{X}^{0'}\Lambda^{-1}\boldsymbol{X}^0}{\boldsymbol{X}^{0'}\Lambda^{-1}\boldsymbol{X}^0 + S^0 + T^2}$$

The orthant probability of a positive definit matrix Σ, denoted by $P\{\Sigma\}$, is the probability that the random vector distributed in $N(0, \Sigma)$ has components all positive.

Theorem The distribution of the statistic $\bar{E}_n^2$ under the null hypothesis is given by the following.

$$\begin{aligned}
\Pr(\bar{E}_n^2 = 1) &= P\{\Lambda\} && \text{in case when } T^2 \text{ is absent} \\
\Pr(\bar{E}_n^2 = 1) &= 0 && \text{in case when } T^2 \text{ is present} \\
\Pr(\bar{E}_n^2 = 0) &= P\{\Lambda^{-1}\} && \text{in both of the above two}
\end{aligned}$$

and for $0 < a < 1$ we have

$$\Pr(\bar{E}_n^2 \geq a) = \Pr(\bar{E}_n^2 = 1) + \sum_{\phi \subset M \subset P} P\{(\Lambda_{M'})^{-1}\} P\{\Lambda_{M:M'}\} I_a\left(\frac{n(M)}{2}, \frac{p - n(M) + t}{2}\right) \quad (1)$$

When T^2 is absent t should be understood as 0, and

$$I_a\left(\frac{k}{2}, \frac{q}{2}\right) = \frac{1}{B(\frac{k}{2}, \frac{q}{2})} \int_0^a x^{\frac{k}{2}-1}(1-x)^{\frac{q}{2}-1}\, dx \quad (2)$$

Here, $\sum$ denotes the summation for all possible non-empty proper subset M of $P = (1, 2, \ldots k)$, $n(M)$ is the number of elements in subset M, M' is the complement of M, Λ_M is the variance matrix of $\boldsymbol{X}_i, i \in M$, $\Lambda_{M:M'}$ is variance matrix under the condition $\boldsymbol{X}_j = 0, j \notin M$, and $P\{A\}$ is the probability that the random variables distributed in a multivariate $N(\mathbf{0}, A)$, where covariance matrix A are all positive.

4 CASES WHEN LR TEST IS POSSIBLE WITHOUT T^2

The equation (1) in this theorem states when T^2 is absent, the likelihood ratio test with the significance level α is possible only when $P\{\Lambda\} < \alpha$.

This condition is satisfied when $\Lambda = I$, the unit matrix and as $0.05 > 2^{-k}$ when $k > 5$.

Another case when this condition is satisfied is the case of simple order alternative. Let $\boldsymbol{X} = (x_1, x_2, \cdots x_n)$ distributed independently in normal with means $\theta_1, \theta_2, \cdots, \theta_n$. the null hypothesis is the equality of them and the alternative is $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_n$ where at least one of $n-1$ inqualities is strict. This problem can be transformed to our frame work by taking the successive difference, and then the condition is seen to be satisfied as $P\{\Lambda\} = \dfrac{1}{n!}$ and this is less than 0.05 when $n > 4$.

5 HISTORICAL REVIEW AND DISCUSSION

The first paper in this area is said to be due to Kudo[2], and cited in [1] and [6]. The recent paper [9] dealt with the computational algorithms based on the method [8].

The case when T^2 does not exist was first treated in [4] and a table was published for simple order case in [3]. When these works were done, the method of calculating the normal orthand probability was not well developed untill [7], which had made one of the authors, Kudo, coward in developing and publishing the general form.

The recent devopement reported in this meeting [10] is most congratulated.

About the checking the accuracy on (1), two identities are available.

$$P\{\Lambda\} + \sum P\{(\Lambda_{M'})^{-1}\} P\{\Lambda_{M:M'}\} + P\{\Lambda^{-1}\} = 1 \tag{3}$$

$$(-1)^n P\{\Lambda\} + \sum_{\phi \subset M \subset P} (-1)^{n(M)} P\{(\Lambda_{M'})^{-1}\} P\{\Lambda_{M:M'}\} + P\{\Lambda^{-1}\} = 0 \tag{4}$$

The first one is trivial and second is due to[5]

References

[1] Barlow R. E., Bartholomew. D. J., Bremner, J. M. and Brunk, H. D.(1972).Statistical Inference under Order Restrictions, 2nd ed. New York:, John Wiley.

[2] Kudô, A. (1963). A multivariate analogue of one-sided test.Biometrika, **50**, 403-18.

[3] Kudô, A. and Yao, J.S. (1982). Tables for testing ordered alternatives in an analysis of variance without replications. Biometrika, **69**, 237–8.

[4] Kudô, A., Sasabuchi S. and Choi, J.R. (1981) Test of equality of normal means in the absence of independent estimator of variance. Commun. Statist.-Theor.Meth. **A10(7)**, 659–668.

[5] mcMullen, P. (1975). Non-linear angle-sum relations for polyhedral cones and polytopes. Math. Proc. Camb. Phil. Soc., **78** 247–61

[6] Robertson, T., Wright, F. T. and Dykstra, R. L.(1988). Order Restricted Statistical Inference, New York:, John Wiley.

[7] Sun, H.-J. (1988). A fortran subroutine for computing normal orthant probabilities of dimensions up to nine.Commun. Statist.-Simula., 17, 3, 1097-111

[8] Tarumi, T. and Kudô, A. (1974). An algorithm related to all possible regression and discriminant analysi. Journ. Japan Statist. Soc. 4, 47-50

[9] Yamamoto, Y. Kudo, A. and Ujiie, K. (1997) Computation of the test statistic and the null distribution in the multivariate analogue of the one sided test. *J.Jpn. Soc. Comp. Statistic.* pp.89-97.

[10] 三輪哲久　Tony Hayter 栗木,(2001) 非心象限確率の効率的計算法. 日本統計学会第６９回総会および研究報告会

Asymptotics for the Likelihood Ratio Test under Loss of Identifiability

Rockefeller Univ. Xin Liu

The classical quadratic expansion of the log-likelihood ratio around the true parameter can fail when parameters characterizing the true distribution are not unique. In finite binomial mixture models, a reparameterization can reduce this problem into the likelihood ratio test under nonstandard conditions. In general, this reparameterization technics may not work. To overcome this difficulty, we establish a quadratic approximation of the log-likelihood ratio function in a Hellinger neighborhood of the true density. Then the asymptotic null distribution of the likelihood ratio test statistic can be obtained by maximizing the quadratic form even when there is loss of identifiability in parameters. Testing the number of components in finite mixture models are considered.

Detecting the number of change-points via likelihood ratio test

九州大学 二宮嘉行

1．はじめに

情報量規準を用いて変化点数を決める理論に比べ, 検定に基づいて変化点数を決める理論は発展していない. 特に複数の変化点に関する検定統計量の分布理論に関する論文は少ない. そこで, 独立に分散既知の正規分布に従う系列の平均パラメータが未知の時点でシフトする, という変化点問題の最も基本的な設定で尤度比統計量の漸近理論を扱い, 数値実験でその妥当性を確かめる.

2．結果

x_i $(i=1,\ldots,m)$ が各々独立に正規分布 $N(\theta_i,1)$ に従うとし, 変化点モデルに従うという仮説

$$H_i\ :\ \theta_1=\cdots=\theta_{k_1}\neq\theta_{k_1+1}=\cdots=\theta_{k_2}\neq\cdots\neq\theta_{k_i+1}=\cdots=\theta_m$$

を考える. ただし漸近論のため k_i は m の関数であって $\lim_{m\to\infty}k_i/m=\lambda_i<1$ を満たすものとする.

x_{s+1} から x_t までの平均 $\sum_{i=s+1}^{t}x_i/(t-s)$ を $\bar{x}_s^t$ と記すことにすると, 変化点数 0 対 1, つまり H_0 対 H_1 の検定の対数尤度比統計量は $T_{0:1}\equiv\max_{1\le s\le m}\{s(\bar{x}_0^s)^2+(m-s)(\bar{x}_s^m)^2-m(\bar{x}_0^m)^2\}$ と書け, 次のような結果がある.

定理 1 (Csörgő & Horváth [1] p.23) $\underline{t}(m)\ge 1/m, \bar{t}(m)\ge 1/m$ において

$$\limsup_{m\to\infty} m\{\underline{t}(m)+\bar{t}(m)\}\exp\{-(\log m)^{1-\epsilon^*}\}<\infty \tag{1}$$

がある $0<\epsilon^*\le 1$ について成り立てば, 任意の $0<\epsilon\le\epsilon^*$ に対し, 帰無仮説 H_0 のもとで

$$\left|\{T_{0:1}-\sup_{\underline{t}(m)\le t\le 1-\bar{t}(m)}\left\{\frac{B(t)^2}{t(1-t)}\right\}^{1/2}\right|=o_P\Big(\exp\{-(\log m)^{1-\epsilon}\}\Big)$$

が成り立つ. ここで, $B(t)$ は Brownian ブリッジで, $\mathrm{E}B(t)=0, \mathrm{E}B(s)B(t)=\min(s,t)-st$ を満たす.

注 1 基準化した Brownian ブリッジの最大値の裾確率に関しては評価式があり, Csörgő & Horváth [1] は $\underline{t}(m)=\bar{t}(m)=(\log m)^{3/2}/m$ としてそれを用いることを提案している.

上で使われている理論を拡張し, 変化点数 n 対 $n+1$, つまり H_n 対 H_{n+1} の検定の対数尤度比統計量 $T_{n:n+1}$ に対して用いる.

定理 2 $\underline{t}_h(m)\ge 1/m, \bar{t}_h(m)\ge 1/m$ $(h=1,\ldots n+1)$ が (1) を満たすならば, 任意の $0<\epsilon<\epsilon^*$ に対し, H_n のもとで

$$\left|\,T_{n:n+1}-\max_{1\le h\le n+1}\left[\sup_{\underline{t}_h(m)\le t\le 1-\bar{t}_h(m)}\left\{\frac{B_h(t)^2}{t(1-t)}\right\}^{1/2}\right]\,\right|=o_P\Big(\exp\{-(\log m)^{1-\epsilon}\}\Big)$$

が成り立つ. ここで $B_h(t)$ $(h=1,\ldots n+1)$ は各々独立な Brownian ブリッジである.

変化点数 0 対 2 の検定を考える前に, エピデミックタイプの変化検知のための尤度比検定に関する結果を述べておく.

定理 3 (Yao [4])

$$H_e\ :\ \theta_1=\cdots=\theta_{k_1}=\theta_{k_2+1}=\cdots=\theta_m\neq\theta_{k_1+1}=\cdots=\theta_{k_2}$$

とし, この変化点の位置に関して $t_0m\le k_2-k_1\le(1-t_1)m$ という制約を入れる (t_0, t_1 は小さい正の定数). この制約の下での H_0 対 H_e の検定の対数尤度比統計量 (一般化尤度比と呼ばれる) を $T_{0:e}$ とし, また $V[\mu]=2\mu^{-2}\exp\{-2\sum_{n=1}^{\infty}n^{-1}\Phi(-|\mu|n^{1/2}/2)\}$ とすれば, 帰無仮説 H_0 のもとで

$$\mathrm{P}(T_{0:e}>m\xi_0)\ \overset{m\to\infty}{\longrightarrow}\ \frac{1}{4\sqrt{2\pi}}(m\xi_0)^{3/2}\exp\Big(-\frac{m\xi_0}{2}\Big)\int_{t_0}^{t_1}\frac{1}{(1-t)t^2}V\Big[\Big\{\frac{\xi_0}{t(1-t)}\Big\}^{1/2}\Big]^2\mathrm{d}t$$

が成り立つ.

このように閾値 $\xi = m\xi_0$ を越える確率の m に関する漸近展開の主項を与えることは，裾確率の大偏差近似と呼ばれる．ここで用いられる Siegmund [2] や Yao [3] の手法を拡張すると，H_0 対 H_2 の一般化尤度比に関する以下の結果を導ける．

定理 4 H_2 の変化点の位置に関して $k_1 \geq t_0 m$, $k_2 - k_1 \geq t_1 m$, $m - k_2 \geq t_2 m$ という制約を入れる (t_0, t_1, t_2 は小さい正の定数)．この制約の下での H_0 対 H_2 の検定の対数尤度比統計量を $T_{0:2}$ とすれば，帰無仮説 H_0 のもとで

$$\mathrm{P}(T_{0:2} > m\xi_0) \overset{m\to\infty}{\longrightarrow} \frac{1}{8\pi} m^2 \exp\Big(-\frac{m\xi_0}{2}\Big) \int_{t_0}^{1-t_2-t_1} \mathrm{d}s \int_{s+t_1}^{1-t_2} \mathrm{d}t \int_{-\sqrt{\xi_0(1-t)/t}}^{\sqrt{\xi_0(1-t)/t}} \mathrm{d}y$$
$$\Big[\sqrt{\frac{t}{1-t}}\Big(\xi_0 - \frac{t}{1-t}y^2\Big)^{-1/2} \Big\{(\nu_{1+}-\nu_{2-})^2(\nu_{2-}-\nu_3)^2 V[\nu_{1+}-\nu_{2-}]$$
$$V[\nu_{2-}-\nu_3] + (\nu_{1-}-\nu_{2+})^2(\nu_{2+}-\nu_3)^2 V[\nu_{1-}-\nu_{2+}]V[\nu_{2+}-\nu_3]\Big\}\Big]$$

が成り立つ．ここで

$$\nu_{1\pm} = y \pm \sqrt{\Big(\xi_0 - \frac{t}{1-t}y^2\Big)\frac{t-s}{st}}, \quad \nu_{2\pm} = y \pm \sqrt{\Big(\xi_0 - \frac{t}{1-t}y^2\Big)\frac{s}{t(t-s)}}, \quad \nu_3 = -\frac{ty}{1-t}.$$

そして H_n 対 H_{n+2} の検定の対数尤度比統計量 $T_{n:n+2}$ に関しては以下の結果を導ける．

定理 5 $T_{j'}^{[0:2]}(m_{j'})$ $(j' = 1, \ldots, n+1)$ を $T_{0:2}$ のコピー（データ数 $m_{j'}$, 各々独立）とすれば，H_n のもとで

$$T_{n:n+2} \sim \max\Big[\max_{1\leq j<h\leq n+1}\Big\{\sup\frac{B_j(t)^2}{t(1-t)} + \sup\frac{B_h(t)^2}{t(1-t)}\Big\}, \max_{1\leq j'\leq n+1} T_{j'}^{[0:2]}(k_{j'}^* - k_{j'-1}^*)\Big] \quad (2)$$

が成り立つ．ここで $k_1^*, \ldots, k_n^*$ は真の変化点であり，また，$k_0^* = 0$, $k_{n+1}^* = m$ としている．

3．数値実験と考察

上で得られた結果で必要とされる定数に対して適当な値を定め，モンテカルロシミュレーションによる評価と比較した．

定理 2 においては，注 1 を考慮し，

$$\underline{t}_j(m) = \bar{t}_j(m) = \{\log(\hat{k}_j - \hat{k}_{j-1})^{3/2}\}/(\hat{k}_j - \hat{k}_{j-1}) \quad (3)$$

を用いることを提案した．ここで $\hat{k}_1, \ldots, \hat{k}_n$ は H_n の下での変化点の最尤推定量であり，その一致性が成り立つことよりこれらは条件 (1) を満たす．そして，変化点数 n に対するデータ数 m の割合が小さすぎなければ，漸近論がうまく働くことを確認した．

定理 4 においては，$t_0 = t_1 = t_2 \geq 0.07$ ならば一般化尤度比の大偏差近似は精度が高いこと，そして，変化点に制約のない場合の尤度比統計量に関しては $t_0 = t_1 = t_2 = 1/m$ を用いるのが自然であるものの，これは m が大きいときによい漸近評価を与えないことを確認した．

定理 5 においては，(2) の $T_{j'}^{[0:2]}(k_{j'}^* - k_{j'-1}^*)$ のかわりに $T_{j'}^{[0:2]}(\hat{k}_{j'} - \hat{k}_{j'-1})$ を用いることと (3) とを提案した．やはり $\hat{k}$ の一致性より，この提案は定理 5 を満たす．$B_j(t)$ と $T_{j'}^{[0:2]}$ は独立ではないが相関は正であるため，保守的な検定の構成を目指し，$B_j(t)$ と $T_{j'}^{[0:2]}$ は独立であるものとして裾確率を近似した．そしてこれは保守的すぎる結果を与えないことを確認した．

参考文献

[1] M. Csörgő and L. Horváth. *Limit Theorems in Change-Point Analysis.* John Wiley & Sons, 1996.

[2] D. Siegmund. Approximate tail probabilities for the maxima of some random fields. *Ann. Probab.*, 16:487–501, 1988.

[3] Q. Yao. Boundary-crossing probabilities of some random fields related to likelihood ratio tests for epidemic alternatives. *J. Appl. Probab.*, 30:52–65, 1993.

[4] Q. Yao. Tests for change-points with epidemic alternatives. *Biometrika*, 80:179–191, 1993.

Selection among the k Bernoulli Trials – Fixed Sample case

明治学院大・国際　竹内 啓

N 個の対象に対して k 種の処理のうちのいずれか適用するとする．結果 X は 0 か 1 のいずれかであり，第 i 処理を適用したとき $P(X=1)=p_i=1-q_i$ となるとする．目的は $E(\sum_{i=1}^{N}X_i)$ を最大にするように処理をえらぶことである．X_i に対する処理の選択は $X_1,\cdots,X_{i-1}$ にのみ依存するとすれば，

$$E\left(\sum_{i=1}^{N}X_i\right)=\sum_{i=1}^{k}p_iE(N_i)=\sum_{i=1}^{k}p_i\nu_i$$

となる．N_i は第 i 処理を適用する回数である．

$\max p_i=p^*$ と置けば regret は

$$R=\sum_{i=1}^{k}(p^*-p_i)\nu_i$$

と定義される．

Fixed sample rule においては最初 kn 個の対象について，k 種の処理を n 個ずつについて施し，残りの $N-kn$ 個については，最も成功の多かったものを施す (タイがあったときはランダムに定める)．最初の n 個についての成功の数を $Y_1,\cdots,Y_k$ とすると，これは 2 項分布 $B(n,p_i)$ に従う．そうすると

$$R=n\sum_{i=1}^{k}(p^*-p_i)+(N-kn)\sum(p^*-p_i)P(Y_i=\max_j Y_j)$$

と表される (タイの場合を考慮すればやや複雑になる)．

$$P(Y_i=\max Y_j)=P(Y_i\geq Y_1,Y_i\geq Y_2,\cdots)$$

の評価に大偏差近似を用いる．$Y_i-Y_1=Z_1,\cdots,Y_i-Y_k=Z_{k-1}$ (Y_i-Y_i を除く) とすると

$$P(Y_i=\max Y_j)=P(Z_1\geq 0,\cdots,Z_{k-1}\geq 0)$$

となる．前稿の方法を用いれば

$$P(Z_1=\cdots=Z_{k-1}=0)=\frac{1}{(2\pi)^{\frac{k-1}{2}}\sqrt{k\bar{P}\bar{Q}}}(\bar{p}+\bar{q})^{nk}$$

$$\bar{p}=\left(\prod p_i\right)^{1/k},\quad \bar{q}=\left(\prod q_i\right)^{1/k},\quad \bar{P}=\bar{p}/(\bar{p}+\bar{q})=1-\bar{Q}$$

となり，$e^{-\hat{\theta}_j}=p_j/\bar{p}$ となるから $p_j<\bar{p}$ $(j=1,\cdots,k,j\neq i)$ ならば

$$P(Z_1\geq 0,\cdots,Z_{k-1}\geq 0)=\frac{(\bar{p}+\bar{q})^{kn}}{(2\pi)^{\frac{k-1}{2}}\sqrt{k\bar{P}\bar{Q}}}\sum\left(\frac{\bar{p}}{\bar{p}-p_i}\right)$$

となる．この近似から，実は

$$R\simeq n\sum(p^*-p_i)+(N-kn)\sum(p^*-p_i)P(Y_i\geq Y_{i^*}).$$

ただし，$p_{i^*}=\max p_i$ となるので，n についての近似的な minimax 解を容易に求めることができる．

Asymptotic properties of estimators for small diffusions

SEI Tomonari* KOMAKI Fumiyasu†

Abstract

The second-order efficiency of bias-corrected estimators for diffusion processes with small noise is investigated from the viewpoint of information geometry. A bias-corrected estimator is second-order efficient if and only if the ancillary manifolds are orthogonal to the model and their embedding m-curvatures vanish. In particular, the maximum likelihood estimator is second-order efficient. It is essential that the statistics appeared in the expansions of estimators have asymptotic normality.

1. Diffusion processes with small noise and information geometry

Suppose that $X^\epsilon = \{X_t^\epsilon \mid t \in [0,T]\} \in \mathcal{C}_T$, where $\mathcal{C}_T$ is the set of continuous functions from $[0,T]$ to $\mathbf{R}$, is a diffusion process with small noise parameter $\epsilon \in [0,1]$

$$\mathrm{d}X_t^\epsilon = \mu(X_t^\epsilon, u)\mathrm{d}t + \epsilon \mathrm{d}W_t, \quad X_0^\epsilon = x_0,$$

where $W = \{W_t\}$ is a 1-dimensional standard Brownian motion, $\mu = \mu(\cdot,\cdot)$ is a smooth function and u is a m-dimensional unknown parameter in the parameter space $U \subset \mathbf{R}^m$, and the initial value x_0 is a constant independent of u. The log likelihood function ℓ_ϵ is given by

$$\ell_\epsilon(X,u) = \frac{1}{\epsilon^2}\int_0^T \mu \mathrm{d}X_t - \frac{1}{2\epsilon^2}\int_0^T \mu^2 \mathrm{d}t.$$

We denote this model by $\mathcal{P} = \{\ell_\epsilon(\cdot,u) | u \in U\}$ and introduce some geometrical notations as follows. See Amari [1] for details about information geometry. A tangent space of $\mathcal{P}$ at $u \in U$ is defined by $\mathcal{T}_u = \mathrm{span}\{\partial_a \ell_\epsilon(\cdot,u)\}$, where $\partial_a = \partial/\partial u^a$ for $a = 1,\cdots,m$. The Fisher information metric $\boldsymbol{g}_{ab}$ and e-, m-connection coefficients $\overset{\mathrm{e}}{\boldsymbol{\Gamma}}_{abc}$, $\overset{\mathrm{m}}{\boldsymbol{\Gamma}}_{abc}$ on $\mathcal{P}$ are defined by

$$\begin{aligned}
\boldsymbol{g}_{ab} &= \mathrm{E}_u[\partial_a \ell_\epsilon \partial_b \ell_\epsilon], \\
\overset{\mathrm{e}}{\boldsymbol{\Gamma}}_{abc} &= \mathrm{E}_u[\partial_a \partial_b \ell_\epsilon \partial_c \ell_\epsilon], \\
\overset{\mathrm{m}}{\boldsymbol{\Gamma}}_{abc} &= \overset{\mathrm{e}}{\boldsymbol{\Gamma}}_{abc} + \mathrm{E}_u[\partial_a \ell_\epsilon \partial_b \ell_\epsilon \partial_c \ell_\epsilon],
\end{aligned}$$

respectively. Since all of these quantities are order of ϵ^{-2}, we put $g_{ab} = \epsilon^2 \mathbf{g}_{ab}$, $\overset{\mathrm{e}}{\Gamma}_{abc} = \epsilon^2 \overset{\mathrm{e}}{\boldsymbol{\Gamma}}_{abc}$ and $\overset{\mathrm{m}}{\Gamma}_{abc} = \epsilon^2 \overset{\mathrm{m}}{\boldsymbol{\Gamma}}_{abc}$. The explicit expressions for these quantities have been obtained (See Sei and Komaki [6]). In the following, we use Einstein's summation convention.

2. The asymptotic expansion for the MLE

The maximum likelihood estimator (MLE) $\hat{u}$ with respect to an observed process X^ϵ has an asymptotic expansion:

$$\begin{aligned}
\hat{u}^a \simeq\ & u^a + \epsilon g^{ab}\tilde{y}_b + \epsilon^2 \Big(g^{ab} g^{cd} \tilde{y}_{bc} \tilde{y}_d \\
& - \frac{1}{2} \overset{\mathrm{m}}{\Gamma}_{cd}{}^a g^{c\epsilon} g^{df} \tilde{y}_\epsilon \tilde{y}_f \Big),
\end{aligned}$$

where

$$\begin{aligned}
\tilde{y}_a &= \epsilon \partial_a \ell_\epsilon, \\
\tilde{y}_{ab} &= \epsilon \partial_a \partial_b \ell_\epsilon - \mathrm{E}_u[\epsilon \partial_a \partial_b \ell_\epsilon] - \overset{\mathrm{e}}{\Gamma}_{ab}{}^c \tilde{y}_c,
\end{aligned}$$

(g^{ab}) is the inverse matrix of (g_{ab}) and the symbol $\simeq$ means asymptotic equivalence. See Kutoyants [5] and Yoshida [7] for the validity of asymptotic expansions related to the MLE.

Using the expansion and asymptotic normality of $\tilde{y}_a$ and $\tilde{y}_{ab}$, we obtain the expansion of the mean square error (MSE) of the bias-corrected MLE $\hat{u}^*$. See Efron [2] and Amari [1] for the concept of bias correction.

Proposition 2.1 *The MSE of the bias-corrected MLE is given by*

$$\mathrm{E}_u[(\hat{u}^{*a} - u^a)(\hat{u}^{*b} - u^b)] \simeq \epsilon^2 g^{ab} + \frac{\epsilon^4}{2} C^{2ab},$$

*Department of Mathematical Engineering and Information Physics, School of Engineering, University of Tokyo.

†Department of Mathematical Informatics, School of Information Science and Technology, University of Tokyo.

where

$$\begin{aligned} C^{2ab} &= (\overset{\mathrm{m}}{\Gamma})^{2ab} + 2(\overset{\mathrm{e}}{\mathrm{H}}_{\mathcal{P}})^{2ab}, \\ (\overset{\mathrm{m}}{\Gamma})^{2ab} &= \overset{\mathrm{m}}{\Gamma}_{cd}{}^{a}\overset{\mathrm{m}}{\Gamma}_{ef}{}^{b} g^{ce} g^{df}, \\ (\overset{\mathrm{e}}{\mathrm{H}}_{\mathcal{P}})^{2ab} &= M_{cdef} g^{ac} g^{be} g^{df}, \\ M_{cdef} &= \mathrm{E}_u[\tilde{y}_{cd}\tilde{y}_{ef}]. \end{aligned}$$

3. The asymptotic expansions for Fisher-consistent estimators

In this section, we assume that the model $\mathcal{P}$ is a curved exponential family embedded in a full exponential family $\mathcal{E} = \{\ell_\epsilon(\cdot,\theta)\}$ defined by

$$\ell_\epsilon(X,\theta) = \epsilon^{-2}(\theta^i s_i(X) - \Psi(\theta,\epsilon)),$$

where θ is the n-dimensional natural parameter, $s(X)$ is the sufficient statistic corresponding to θ and $\Psi(\theta,\epsilon)$ is the potential function. The quantity ϵ plays a role as the dispersion parameter (Jørgensen [3]). See Küchler and Sørensen [4] for details about curved exponential families of stochastic processes. This assumption holds if the drift coefficient is in a linear form.

Let η be the expectation parameter: $\eta_i = \eta_i(u,\epsilon) = \mathrm{E}_u[s_i(X^\epsilon)]$, which depends on ϵ. We suppose that the asymptotic normality

$$\tilde{\eta}_i = \epsilon^{-1}(s_i(X) - \eta_i) \to N(0,(g_{ij})) \quad (1)$$

holds.

Let us consider an estimator expressed by $\hat{u} = T(\hat{\eta})$, where T is a smooth function from R^n to U. We assume that $\hat{u}$ is Fisher consistent, i.e., $u^a = T^a(\eta(u,0))$ holds for all $u \in U$. We define the ancillary manifold $\mathcal{A}_u$ by $\mathcal{A}_u = T^{-1}(\{u\})$. The notations about the ancillary manifolds follow Amari[1].

The estimator $\hat{u}$ is expanded as

$$\hat{u}^a \simeq u^a + \epsilon\{\partial^i T^a\}\tilde{\eta}_i + \frac{\epsilon^2}{2}\{\partial^i\partial^j T^a\}\tilde{\eta}_i\tilde{\eta}_j.$$

Here, $\partial^i T^a(\eta)$ $(a = 1,\cdots,m)$ are interpreted as vectors normal to the ancillary manifold $\mathcal{A}_u$. It can be shown that the estimator $\hat{u}$ is first-order efficient if and only if $\mathcal{A}_u$ and $\mathcal{T}_u$ are asymptotically orthogonal, i.e., $\partial^i T^a \simeq g^{ab}\partial_b\theta^i$.

From now on, we assume that $\hat{u}$ is first-order efficient and discuss second-order efficiency of $\hat{u}$. By using asymptotic normality (1) of $\tilde{\eta}$ and a relation $\partial^i\partial^j T^a = -A^{i\beta}A^{j\gamma}\overset{\mathrm{m}}{\Gamma}_{\beta\gamma}{}^{a}$, where $A^{i\beta} = \partial^i w^\beta$, the MSE of the bias-corrected estimator $\hat{u}^*$ is represented as follows.

Theorem 3.1 *The MSE of a bias-corrected first-order efficient estimator is*

$$\mathrm{E}_u[(\hat{u}^{*a} - u^a)(\hat{u}^{*b} - u^b)] \simeq \epsilon^2 g^{ab} + \frac{\epsilon^4}{2}C^{2ab}$$

where

$$\begin{aligned} C^{2ab} &= (\overset{\mathrm{m}}{\Gamma})^{2ab} + 2(\overset{\mathrm{e}}{\mathrm{H}}_{\mathcal{P}})^{2ab} + (\overset{\mathrm{m}}{\mathrm{H}}_{\mathcal{A}})^{2ab}, \\ (\overset{\mathrm{m}}{\mathrm{H}}_{\mathcal{A}})^{2ab} &= \overset{\mathrm{m}}{\Gamma}_{\kappa\lambda}{}^{a}\overset{\mathrm{m}}{\Gamma}_{\mu\nu}{}^{b} g^{\kappa\mu} g^{\lambda\nu}, \end{aligned}$$

and $(\overset{\mathrm{m}}{\Gamma})^{2ab}$ *and* $(\overset{\mathrm{e}}{\mathrm{H}}_{\mathcal{P}})^{2ab}$ *are given in proposition 2.1.*

The three terms in C^{2ab} are all non-negative and $(\overset{\mathrm{m}}{\Gamma})^{2ab}$ and $(\overset{\mathrm{e}}{\mathrm{H}}_{\mathcal{P}})^{2ab}$ are independent of the estimator. Therefore, $\hat{u}^*$ is second-order efficient if and only if $(\overset{\mathrm{m}}{\mathrm{H}}_{\mathcal{A}})^{2ab} \to 0$ as $\epsilon \to 0$. The quantity $(\overset{\mathrm{m}}{\mathrm{H}}_{\mathcal{A}})^{2ab}$ means square of the embedding m-curvature of $\mathcal{A}_u$. In particular, the bias-corrected MLE is second-order efficient.

References

[1] Amari, S. (1985). *Differential-Geometrical Methods in Statistics*, Lecture Notes in Statistics 28, Springer-Verlag.

[2] Efron, B. (1975). *Annals of Statistics*, **3**, 1189-1242.

[3] Jørgensen, B. (1992). *International Statistical Review*, **60**, 5-20.

[4] Küchler, U. and Sørensen, M. (1997). *Exponential Families of Stochastic Processes*, Springer Series in Statistics, Springer-Verlag New York, Inc..

[5] Kutoyants, Yu. A. (1984). *Theory of Probability and its Applications* , **29**, 465-477.

[6] Sei, T. and Komaki, F. (2002). in preparation.

[7] Yoshida, N. (1992). *Probability Theory and Related Fields*, **92**, 275-311.

進化速度の確率変動モデル：ゲノム進化と相同性検索

東京大学・農学生命科学　岸野洋久
ノースカロライナ州立大学・バイオインフォーマティクス　Jeffrey L. Thorne

1　序

数多くあるゲノムの変異の中から、生物の適応進化に本質的に結びついたものを探し出す努力が多くの生物科学者によって積み重ねられてきた。現在のところ、ゲノムレベルでの適応進化の爪痕を検出する際の第一の探索的アプローチとして系統間で分子進化速度を比較する方法が有力視されている。重複遺伝子の運命([6, 4])、歯類の共生と分子進化速度の加速化([3])など、90年代半ばから興味深い研究が次々に出てきた。

筆者らはこれまで、分子進化速度が変動する様子を、事前にグループ分けするなど強い制約を置くことなく柔軟に推定する階層モデルを提案した([5])。そして、分子時計を仮定した解析との対比において、分岐年代推定の頑健性を調べた([2])。

2　分子系統樹の尤度と進化速度の確率変動モデル

マルコフ過程で記述される分子進化の推移速度行列をモデル化することにより、分子系統樹の尤度が表現される。相同な s 本の配列を比較して系統関係を推定する場合を考える。配列の長さを n とすると、系統樹 T の対数尤度は

$$l(\boldsymbol{\theta}|\mathbf{X}) = \sum_{h=1}^{n} \log f(\mathbf{X}_h|\boldsymbol{\theta}) \tag{1}$$

と表わされる。ここで $\boldsymbol{\theta}_i$ は進化のプロセスを規定するパラメータである。配列の変化の統計モデルとしてはマルコフ過程によるモデル化が妥当する。分岐後それぞれの種の配列は独立に進化すると仮定すると、$f(\mathbf{X}|\boldsymbol{\theta})$ は

$$f(\mathbf{X}|\boldsymbol{\theta}) = \sum_{Z_{i_0}} \pi_{Z_{i_0}} \prod_{j \in node(T)\setminus i_0} \sum_{Z_j} P_{Z_{anc(j)},Z_j}(t_{anc(j),j}) \tag{2}$$

と簡単に表される。ここで、$node(T)$ は系統樹 T の節を表し、i_0 はその根である。無根系統樹の場合には、任意の節を指定する。$anc(j)$ は j に隣接する祖先となる節である。$P_{xy}(t)$ は時間 t を経た推移確率である。

進化速度が変動する背景要因として、選択圧の変化を中心とした環境変動、集団の大きさの変動、世代の長さの変動などが考えられる。これらはいずれも自己相関を持って変動することが予想される。そこで、事前分布として速度 $r(t)$ の対数をとったものが簡単な拡散過程に従うとする。すなわち、$\tilde{r}(t) = \log r(t)$ は正規マルコフ過程で、任意の2時点 t, s $(t > s)$ に対して

$$\begin{aligned} E\left[\tilde{r}(t)|\tilde{r}(s)\right] &= \tilde{r}(s) - \frac{\nu}{2}(t-s) \\ V\left[\tilde{r}(t)|\tilde{r}(s)\right] &= \nu(t-s) \end{aligned}$$

を仮定する。移流項は進化速度の期待値が傾向的に変動させないためにつけたものである。分岐後、速度は2系統で独立に変化するとする。各枝の平均速度は、それをはさむ2つの節における速度の平均で近似する。分岐年代を所与とした系統樹の節における速度の対数は、多変量対数正規分布に従う。

分布を規定する超パラメータとして平均速度 μ および拡散係数 ν を持つ。これら2つの超パラメータは独立なガンマ分布に従うとする。時間スケールを規格化し、μ の期待値は1、標準偏差は0.5に設定する。また、ν については強い制約とならないよう、平均、分散ともに2.0に設定する。複数の遺伝子を扱う場合には、独立な事前分布を導入する。

3 複数遺伝子の解析と共進化の検出

遺伝子間で独立な速度変動の事前分布を導入し、分岐年代を共有させた複数の遺伝子の速度変化階層モデルを検討する。シミュレーションを通じて、速度変動の大きさを個々の遺伝子について精度良く推定するためには数多くの配列が必要で、分岐年代の推定については、数多くの遺伝子を解析するときにこの影響がより強く現れることがわかった。ただし、RNA ウイルスのように異なる時点から採られた配列には、分子進化速度に関する情報が多く含まれている。

ところで、遺伝子間で独立を仮定した事前分布に対する速度の事後平均を遺伝子間で比較することにより、共進化を検出することができる。ここで、速度の系統間の系列相関があることに注意が必要である。2 遺伝子における各節での速度の事後平均 $\overline{r}_{mj}$ $(m = 1, 2,\ j \in node(T))$ を求め、順位相関を取る。無相関の帰無仮説の下での分布は、速度変化の確率変動と推定誤差による不確実性を踏まえていなければならない。ここでは、速度変化の方向を無作為化することにより分布を得る。すなわち、帰無仮説の下での各節における速度の事後平均 $\overline{r}'_{mj}$ を

$$\begin{aligned} \overline{r}'_{m,i_0} &= \overline{r}_{m,i_0} \\ \overline{r}'_{m,j} &= \overline{r}'_{m,ranc(j)} + W_{mj}\left(\overline{r}_{m,j} - \overline{r}_{m,ranc(j)}\right) \qquad (j \in node(T) \setminus i_0) \end{aligned}$$

により生成する。W_{mj} は 1, -1 をそれぞれ確率 $\frac{1}{2}$ でとる互いに独立な確率変数である。

4 比較ゲノムとホモロジー検索

ゲノムはミクロレベル・マクロレベルで変化し続けているが、その大枠はかなりの程度で保存されていることがわかって来た ([1])。進化の歴史に基づくゲノムの相関関係は、重要な遺伝子を探索するときに他の生物で得られた知見を利用できることを示唆している。ただ詳細なスケールでは、マーカーや遺伝子の相対的な位置関係に結構入れ替わりがあるという証拠が得られつつある。

比較ゲノムにより種を跨いだゲノムの対応関係を詳細に追って行くときには、注意が必要である。相同性検索はひとつの統計的推定作業であり、誤差が付きまとう。また、ある時点で遺伝子重複が起こり、離れた場所にこの遺伝子のコピーが生じたような場合も見かけ上遺伝子が移動したように映ってしまうことがある。ゲノムの変異と多様化に対する多くの人の理解が深まるにつれ、相同性検索における不確実性を考慮に入れた解析の重要性が今後認識されてくるであろう。

参考文献

[1] Gale, M. D. and Devos, K. M. *Science.* **282**: 656 - 659 (1998)

[2] Kishino, H., Thorne, J. L., and Bruno, W. J. *Mol. Biol. and Evol*, **18**: 352-361 (2001)

[3] Lutzoni, F. and Pagel, M. *Proceedings of National Academy of Sciences, USA.* **94**: 11422-11427 (1997)

[4] Lynch, M. and Conery, J. S. *Science.* **290**: 1151-1155 (2000)

[5] Thorne, J. L., Kishino, H. and Painter, I. S.. *Mol. Biol. Evol.* **15**: 1647-1657 (1998)

[6] Zhang, J., Rosenberg, H. F. and Nei, M. *Proceedings of National Academy of Sciences, USA.* **95**: 3708-3713 (1998)

QTL 解析における非正則性: 遺伝的アルゴリズムによる大域的最適化と閾値

東京大学大学院農学生命科学研究科　中道礼一郎

1　目的

生物学研究において量的形質に関与する遺伝子座（quantitative trait locus, QTL）の解析は、重要な課題である。量的形質は環境など遺伝因子以外の要素によって変動し、比較的効果の小さい複数の遺伝子座によって支配されているため、その解析には統計遺伝学的手法が必要である。現在最も一般的なのは、連鎖地図上の全ての位置で QTL の存在を尤度により検定する手法である[3]。しかし、QTL の数は事前に知ることができないため、可能な全ての QTL の数と位置の組み合わせに対して検定を行うのは現実的ではなく、効率的な探索法が必要である。また、モデルに取り込む QTL の数に対してペナルティーを課すため、検定統計量に対し適切な閾値を定める必要もある。本報告では、F_2 集団と無作為交配集団を対象に、情報量規準(Akaike's information criterion, AIC)を評価関数とする遺伝的アルゴリズム(genetic algorithm, GA)を用いて、最適な QTL モデルを選択する手法を提案する。

2　複数 QTL の遺伝モデルと尤度

一般的な線形モデルのもとで、ある生物個体の表現型値 Y が次のように表される。

$$Y=\mu+\sum_{m=1}^{M}g_m+e$$

ここで、μ は遺伝子型によらない定数である。e は環境効果による残差であり平均 0、分散一定の正規分布に従うと仮定する。M は QTL の数で、g_m（$m=1,2,\cdots,M$）は m 番目の QTL の遺伝効果である。簡単のため QTL 間に遺伝子間相互作用（epistasis）は無いものとしている。

自殖や近交によって純系の親系統を作ることが容易な生物では、2 つの純系親を交配して得られる分離集団を用いて検定を行う。F_2 集団の場合、M 個の QTL に対する尤度 L が以下のように得られる。

$$L=\prod_{j=1}^{N}\sum_{k=1}^{3^M}\left\{P_{j,k}\cdot\phi_j\left(\sum_{m=1}^{M}g_{m,k}\right)\right\}$$

ここで、N は F_2 個体の数で、$P_{j,k}$ は j 番目（j=1,2,$\cdots$,N）の F_2 個体の m 番目（m=1,2,$\cdots$,M）の QTL が、隣接マーカーの遺伝子型にたいして k 番目（k=1,2,$\cdots$,3^M）の分離パターンに該当する条件つき確率である。この確率は QTL とマーカーの位置から組換価によって求められる。ϕ_j は QTL 遺伝子型に対する量的形質の条件付分布である。前述のように量的形質の残差は正規分布を仮定するので以下のように表される。

$$\phi_j(G)=\frac{1}{\sqrt{2\mathrm{p}\sigma^2}}\exp\left\{-\frac{(y_j-\mu-G)^2}{2\sigma^2}\right\}$$

ここで σ^2 は残差分散であり、$g_{m,k}$ は m 番目の QTL が k 番目の分離パターンに該当するときの遺伝効果である。

純系の親系統を作るのが困難な生物では交配実験のデザインがやや複雑になる。現在、我々が採用しているデザインでは、調査する集団から無作為に N_p 個体抽出して親個体とする。この親個体間で N_p (N_p-1) / 2 組の交配を行い、交配ごとに N_{F1} 個体の F_1 集団を得る。このとき、M 個の QTL に対する尤度 L は以下のようになる。

$$L=\prod_{i=1}^{N_P}\left\{\phi_i\left(\sum_{m=1}^{M} g^{P}_{m,i}\right)\right\}\times \prod_{l=1}^{\frac{N_P(N_P-1)}{2}}\prod_{j=1}^{N_{F1}}\left\{\sum_{k=1}^{4^M}\left(P_{j,k}\times\phi_j\left(\sum_{m=1}^{M} g^{F1}_{m,k}\right)\right)\right\}$$

ここで、$g^{P}_{m,i}$はi番目（i=1,2,…,N_P）の親個体のm番目（m=1,2,…,M）のQTLの遺伝効果で、$P_{j,k}$はj番目(j=1,2,…,N_{F1})のF_1個体のQTLが、隣接マーカーの遺伝子型にたいしてk番目(k=1,2,…,4^M）の分離パターンに該当する条件つき確率であり、$g^{F1}{}_{m,k}$はそのときの遺伝効果である。

3　GAによる大域探索とAICによるペナルティー

QTLの数と位置を仮定すれば、その仮定を尤度によって評価することが可能である。そこで、尤度を最大にするQTLの数と位置の組み合わせを遺伝的アルゴリズム (genetic algorithm, GA)によって探索する。今回設計したGAでは、GA遺伝子型はQTLの数と位置である。M個のQTLがあるとき、これは$2M$+1個の要素の数列、M, (c_1,l_1), (c_2,l_2) ... (c_M,l_M) であらわされる。ここで、c_m（m=1,2,…,M）はm番目のQTLの存在する染色体の番号であり、l_mはm番目のQTLの染色体c_m上での位置である。l_mは、染色体c_mの端からの距離（cM）によって表される。このGA遺伝子型に対するGA適応度には、これらのM個のQTLに対する尤度を用いるが、QTLの存在しない位置にゴーストを検出することを抑えるため、情報量規準（AIC）を用いることにより、QTL数にペナルティーを課す。

GA fitness = $-2\ln L+2$(パラメータ数)

このようなGA個体の集団をはじめはランダムに生成し、適応度に応じた淘汰と交配と突然変異を繰り返して新たなGA個体を生成し、より高い尤度を持つQTLの数と位置の組み合わせを探索する。交配では、適応度に応じて2つのGA個体を取り出し、それらのGA遺伝子型に含まれるQTLの位置から、無作為にQTLの位置を選択して、新しいGA個体の遺伝子型を生成する。また、突然変異では、GA遺伝子型に含まれるQTLの位置を無作為に変更する。これには、QTL数の増減を含む大幅な変異と、QTLの位置をわずかにずらすだけの小幅な変異をくみあわせる。

4　GAによる優れた探索性能と今後の課題

シミュレーションによって生成した集団を用いて、さまざまな条件でQTL検出実験を行い、従来の手法との比較も行った。実験に用いた個体のマーカー遺伝子型データとQTL遺伝子型データは、Haldaneのモデルに基づいて、遺伝子座間で干渉が無いものとして生成した。量的形質の表現型値は、複数QTLモデルに基づき正規分布に従う乱数として生成した。その結果、様々な条件のもとで、GAは従来のstepwise探索[2]やBayes推定[4][5]より優れた大域探索性能を示した。一方、検出の閾値の設定は悩ましい問題である。農学分野では、QTL解析の主目的は有用形質をもつ系統を選抜し育種を行うことであるため、今回我々は比較的liberalな規準であるAICを採用したが、よりconservativeな、Permutation Test[1]や検定統計量の分布の近似[3]によって閾値を求める手法も広く使われている。複数のQTLを扱う場合には、確実な誤り率のコントロールと計算負荷の軽減の両立が欠かせないが、現時点ではどの手法も完全ではない。

参考文献

[1] Churchill, G.A. and Doerge, R.W. (1994) *Genetics* **138**: 963-971.
[2] Kao, C. H., Zeng, Z.-B. and Teasdale, R. D. (1999) *Genetics* **152**: 1203-1216.
[3] Lander, E. S. and Botstein, D. (1989) *Genetics* **121**: 185-199
[4] Sillanpää, M. J. and Arjas, E. (1999) *Genetics* **148**: 1373-1388
[5] Stephens, D. A. and Fisch, R. D. (1998) *Biometrics* **54**: 1334-1347

Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling

Hidetoshi Shimodaira
The Institute of Statistical Mathematics shimo@ism.ac.jp

1 Introduction

We consider a function taking 0/1-value, denoted $H_0(\mathcal{X}_n)$, of data $\mathcal{X}_n = \{x_1, \ldots, x_n\}$. (Ex. 1) $H_0(\mathcal{X}_n) = 1$ if $\bar{x} = (x_1 + \cdots + x_n)/n \leq 0$, and $H_0(\mathcal{X}_n) = 0$ otherwise. (Ex. 2) $H_0(\mathcal{X}_n) = 1$ if {rabbit, mouse} clade is supported by a cluster analysis of mammalian DNA sequences, and $H_0(\mathcal{X}_n) = 0$ otherwise. Let $H_0(\mathcal{X}_\infty)$ be the limiting value of $H_0(\mathcal{X}_n)$ as $n \to \infty$, and $H_0(\mathcal{X}_\infty)$ is assumed to be what we want know here. However, observing $H_0(\mathcal{X}_n) = 0$ may not necessarily imply $H_0(\mathcal{X}_\infty) = 0$, since randomness of the data $\mathcal{X}_n$ leads to randomness of $H_0(\mathcal{X}_n)$. Therefore we consider a hypothesis testing of the null hypothesis $H_0(\mathcal{X}_\infty) = 1$ against the alternative $H_0(\mathcal{X}_\infty) = 0$, and decide $H_0(\mathcal{X}_\infty) = 0$ when the null hypothesis is rejected. The p-value, denoted $\hat{\alpha}(\mathcal{X}_n)$, of the test takes a real-value between 0 and 1, such that

$$\Pr\{\hat{\alpha}(\mathcal{X}_n) < \alpha\} \leq \alpha, \quad H_0(\mathcal{X}_\infty) = 1, \quad \text{and} \tag{1}$$

$$\Pr\{\hat{\alpha}(\mathcal{X}_n) < \alpha\} \geq \alpha, \quad H_0(\mathcal{X}_\infty) = 0 \tag{2}$$

hold at least approximately.

In practice, $H_0(\mathcal{X}_n)$ is given as a procedure (computer software), and its analytical expression is hard to obtain. Our p-value calculation described below works perfectly even in such a case. Our method accesses the data only through the function $H_0(\mathcal{X}_n)$ and resampling of $\mathcal{X}_n$. The same feature is shared by the naive bootstrap and the double bootstrap. The advantage of our method is that it achieves the same asymptotic order of accuracy as the double bootstrap (third-order accurate), yet it requires only the same order of computation as the naive bootstrap (linear in the number of replicates).

2 Multistep-Multiscale Bootstrap Resampling

Let $\mathcal{X}^*_{n_1} = \{x^*_1, \ldots, x^*_{n_1}\}$ be a replicate dataset obtained by resampling n_1 elements with replacement from the data $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ of sample size n. Usually $n_1 = n$, but we reserve the generality of using any value for n_1. Let $\mathcal{X}^{**}_{n_2} = \{x^{**}_1, \ldots, x^{**}_{n_2}\}$ be a replicate dataset obtained by resampling n_2 elements with replacement from the replicate $\mathcal{X}^*_{n_1}$. We repeat this step again, and $\mathcal{X}^{***}_{n_3} = \{x^{***}_1, \ldots, x^{***}_{n_3}\}$ is a replicate obtained by resampling n_3 elements with replacement from $\mathcal{X}^{**}_{n_2}$. Let $(n_1^{(k)}, n_2^{(k)}, n_3^{(k)})$, $k = 1, \ldots, K$ be K combinations of (n_1, n_2, n_3). For each combination, $(\mathcal{X}^*_{n_1}, \mathcal{X}^{**}_{n_2}, \mathcal{X}^{***}_{n_3})$ is repeatedly generated $B^{(k)}$ times; $C_1^{(k)}$ is the observed frequency of $H_0(\mathcal{X}^*_{n_1}) = 1$, $C_2^{(k)}$ is that of $H_0(\mathcal{X}^{**}_{n_2}) = 1$, and $C_3^{(k)}$ is that of $H_0(\mathcal{X}^{***}_{n_3}) = 1$. The total number of replicates is $3\sum_{k=1}^K B^{(k)}$, rather than $\sum_{k=1}^K (B^{(k)})^3$.

Each $C_3^{(k)}$ is binomially distributed, and the limiting value as $B^{(k)} \to \infty$ is denoted by $C_3^{(k)}/B^{(k)} \to \pi_3(n/n_1^{(k)}, n/n_2^{(k)}, n/n_3^{(k)})$. $\pi_1(n/n_1^{(k)})$ and $\pi_2(n/n_1^{(k)}, n/n_2^{(k)})$ are similarly defined. These functions $\pi_i(\cdot)$ are related to the p-value as described below.

3 Approximately Unbiased Tests

Let us assume the existence of unknown smooth transformation $y = f(\mathcal{X}_n)$ such that p-dimensional random vector $y = (y_1, \ldots, y_p)$ belongs to the exponential family $y \sim \exp(\theta^i y_i - \psi(\theta) - h(y))$. Using the expectation parameter $\eta_i = \partial\psi/\partial\theta^i$, the region of $H_0(\mathcal{X}_\infty) = 1$ is assumed to be expressed as $H_0 = \{\eta = (\eta_1, \ldots, \eta_p) \mid \eta_p \leq -d^{ab}\eta_a\eta_b - e^{abc}\eta_a\eta_b\eta_c\}$, where the indices a, b, c run through $1, \ldots, p-1$. Without loss of generality we assume the potential $\phi(\eta) = \max_\theta\{\theta^i\eta_i - \psi(\theta)\}$ satisfies $\partial\phi/\partial\eta_i|_0 = 0$, $\partial^2\phi/\partial\eta_i\partial\eta_j|_0 = \delta_{ij}$ and the observed data is $y = (0, \ldots, 0, \lambda)$. Then the following holds with error $O(n^{-3/2})$.

$$\begin{aligned}\pi_3(\tau_1^2, \tau_2^2, \tau_3^2) &= 1 - \Phi\Big\{\gamma_1 T_1 \left(1 + \gamma_3 T_2 + 4\gamma_3^2 T_2^2 + \gamma_5 T_3 + \gamma_6 T_4\right) \\ &\qquad -(\gamma_1 T_1)^{-1}\left(\gamma_2 + \gamma_3 T_2 + 7\gamma_3^2 T_2^2 + \gamma_4 T_2 + 3\gamma_5 T_3 + 3\gamma_6 T_4\right)\Big\}, \qquad (3)\end{aligned}$$

where $T_1 = (\tau_1^2 + \tau_2^2 + \tau_3^2)^{-1/2}$, $T_2 = (\tau_1^2\tau_2^2 + \tau_2^2\tau_3^2 + \tau_3^2\tau_1^2)T_1^4$, $T_3 = (\tau_1^2\tau_2^2\tau_3^2 + \tau_2^4\tau_3^2 + \tau_1^4(\tau_2^2 + \tau_3^2))T_1^6$, and $T_4 = (\tau_1^2\tau_2^2\tau_3^2)T_1^6$. The coefficients $\gamma_1, \ldots, \gamma_6$ are defined geometrically by

$$\begin{aligned}
\gamma_1 &= \lambda + \tfrac{1}{3}\lambda^2\phi^{ppp} + \lambda^3\Big\{-\tfrac{1}{8}\phi^{app}\phi^{app} - \tfrac{1}{18}(\phi^{ppp})^2 + \tfrac{1}{8}\phi^{pppp}\Big\}, \\
\gamma_2 &= \lambda\Big\{-d^{aa} - \tfrac{1}{6}\phi^{ppp}\Big\} + \lambda^2\Big\{d^{ab}d^{ab} - \tfrac{1}{2}d^{aa}\phi^{ppp} + \tfrac{1}{8}\phi^{app}\phi^{app} + \tfrac{1}{72}(\phi^{ppp})^2 - \tfrac{1}{24}\phi^{pppp}\Big\}, \\
\gamma_3 &= -\tfrac{1}{6}\lambda\phi^{ppp} + \lambda^2\Big\{\tfrac{1}{4}\phi^{app}\phi^{app} + \tfrac{1}{9}(\phi^{ppp})^2 - \tfrac{1}{8}\phi^{pppp}\Big\}, \\
\gamma_4 &= \lambda^2\Big\{-d^{ab}\phi^{abp} + \tfrac{1}{3}d^{aa}\phi^{ppp} + \tfrac{1}{2}\phi^{abp}\phi^{abp} + \tfrac{1}{2}\phi^{app}\phi^{app} + \tfrac{2}{9}(\phi^{ppp})^2 - \tfrac{1}{4}\phi^{aapp} - \tfrac{1}{6}\phi^{pppp}\Big\}, \\
\gamma_5 &= \lambda^2\Big\{-\tfrac{1}{8}\phi^{app}\phi^{app} - \tfrac{1}{8}(\phi^{ppp})^2 + \tfrac{1}{12}\phi^{pppp}\Big\}, \quad \gamma_6 = \lambda^2\Big\{-\tfrac{1}{8}\phi^{app}\phi^{app} - \tfrac{1}{8}(\phi^{ppp})^2 + \tfrac{1}{24}\phi^{pppp}\Big\},
\end{aligned}$$

where $\phi^{ijk} = \partial^3\phi/\partial\eta_i\partial\eta_j\partial\eta_k|_0$, $\phi^{ijkl} = \partial^4\phi/\partial\eta_i\partial\eta_j\partial\eta_k\partial\eta_l|_0$.

Since $\lambda = O(1)$, $d^{ab} = O(n^{-1/2})$, $\phi^{ijk} = O(n^{-1/2})$, $\phi^{ijkl} = O(n^{-1})$, the coefficients are $\gamma_1 = O(1)$, $\gamma_2, \gamma_3 = O(n^{-1/2})$, $\gamma_4, \gamma_5, \gamma_6 = O(n^{-1})$. The coefficients $\gamma_1, \ldots, \gamma_6$ are estimated by fitting $\pi_3(\tau_1^2, \tau_2^2, \tau_3^2)$ to $C_3^{(k)}$, $k = 1, \ldots, K$ obtained by the multistep-multiscale bootstrap. Using the coefficients, we can calculate the p-value

$$\hat{\alpha}(y) = 1 - \Phi\Big\{\gamma_1\left(1 + \gamma_3 + 4\gamma_3^2 + \gamma_6\right) + \gamma_1^{-1}\left(\gamma_2 + \gamma_3^2/2 + \gamma_4 + \gamma_5\right)\Big\}, \qquad (4)$$

which satisfies (1) and (2) with error $O(n^{-3/2})$. We also obtain some information from $\pi_1(\tau_1^2) = \pi_3(\tau_1^2, 0, 0) = 1 - \Phi(\gamma_1/\tau_1 - (\gamma_2/\gamma_1)\tau_1)$, $\pi_2(\tau_1^2, \tau_2^2) = \pi_3(\tau_1^2, \tau_2^2, 0)$. For example, $\gamma_1, \gamma_2, \gamma_3$ are estimated from $C_2^{(k)}$ so that the p-value is calculated with error $O(n^{-1})$. When y is normally distributed, $\gamma_3 = \cdots = \gamma_6 = 0$ and thus the p-value is calculated with error $O(n^{-3/2})$ from $C_1^{(k)}$. In the normal case, the results of [1] and [2] lead to $\pi_1(\tau_1^2)$ and $\hat{\alpha}(y)$, and the onestep-mutiscale bootstrap is proposed in [3]; the computer software [4] is available from the author.

Bibliography [1] EFRON, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika* **72**, 45–58. [2] EFRON, B. AND TIBSHIRANI, R. (1998). The problem of regions. *Ann. Statist.* **26**, 1687–1718. [3] SHIMODAIRA, H. (2000). Another calculation of the p-value for the problem of regions using the scaled bootstrap resamplings. Technical Report No. 2000-35, Stanford University. [4] SHIMODAIRA, H. AND HASEGAWA, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246–1247.

Large Deviation Approximation of Multivariate Distributions

明治学院大・国際　竹内 啓

$\boldsymbol{X}_i$ $(i=1,\cdots,n,\cdots)$ が i.i.d. 実 p - ベクトル確率変数であるとき

$$\bar{\boldsymbol{X}}_n = \sum_{i=1}^{n} \boldsymbol{X}_i/n$$

と定義し,

$$P(\bar{\boldsymbol{X}}_n \geq \boldsymbol{x}) = Q_n(\boldsymbol{x})$$

の漸近値を考える. $\boldsymbol{X}_i$ の m.g.f.

$$M(\boldsymbol{\theta}) = E\left[\exp(\boldsymbol{\theta}'\boldsymbol{X}_1)\right]$$

が実ベクトル $\boldsymbol{\theta}$ の原点をふくむ開集合と定義されるとき, 大偏差近似が可能になる.

1) $\boldsymbol{X}_i$ の分布が連続な密度関数 $f(\boldsymbol{x})$ を持つとき, 指数型分布族

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = M(\boldsymbol{\theta})^{-1} e^{\boldsymbol{\theta}'\boldsymbol{x}} f(\boldsymbol{x})$$

を定義する. $\bar{X}_n$ の密度関数を $p_n(\boldsymbol{x}|\boldsymbol{\theta})$ と表すと

$$p_n(\boldsymbol{x}|\boldsymbol{\theta}) = M(\boldsymbol{\theta})^{-n} e^{n\boldsymbol{\theta}'\boldsymbol{x}} f_n(\boldsymbol{x}) \quad (f_n(\boldsymbol{x}) = p_n(\boldsymbol{x}|\boldsymbol{0}))$$

となるから

$$f_n(\boldsymbol{x}) = M(\boldsymbol{\theta})^n e^{-n\boldsymbol{\theta}'\boldsymbol{x}} p_n(\boldsymbol{x}|\boldsymbol{\theta}).$$

$\bar{\boldsymbol{X}}_n$ の特性関数は,

$$E\left[\exp(i\boldsymbol{t}'\bar{\boldsymbol{X}}_n)|\boldsymbol{\theta}\right] = M(\boldsymbol{\theta})^{-n} M\left(\boldsymbol{\theta} + \frac{i\boldsymbol{t}}{n}\right)^n$$

となるから

$$p_n(\boldsymbol{x}|\boldsymbol{\theta}) = (2\pi)^{-p} \int M\left(\boldsymbol{\theta} + \frac{i\boldsymbol{t}}{n}\right)^n M(\boldsymbol{\theta})^{-n} e^{-i\boldsymbol{t}'\boldsymbol{x}} d\boldsymbol{t}$$

$$f_n(\boldsymbol{x}) = (2\pi)^{-p} M(\boldsymbol{\theta})^n e^{-n\boldsymbol{\theta}'\boldsymbol{x}} \int M\left(\boldsymbol{\theta} + \frac{it}{n}\right)^n M(\boldsymbol{\theta})^{-n} e^{-i\boldsymbol{t}'\boldsymbol{x}} d\boldsymbol{t}$$

を得る. ここで

$$K(\boldsymbol{\theta}) = \log M(\boldsymbol{\theta}),$$

$$K_i(\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_i} K(\boldsymbol{\theta}), \qquad K_{ij}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial\theta_i \partial\theta_j} K(\boldsymbol{\theta}), \qquad \text{etc.}$$

と表し, $\hat{\boldsymbol{\theta}}$ を

$$K_i(\hat{\boldsymbol{\theta}}) = \boldsymbol{x}_i \qquad (i=1,\cdots,p)$$

をみたすとすれば

$$M\left(\hat{\boldsymbol{\theta}} + \frac{i\boldsymbol{t}}{n}\right)^n M(\hat{\boldsymbol{\theta}})^{-n} = \exp\left(-\frac{1}{2n}\sum\sum K_{ij}(\hat{\boldsymbol{\theta}}) t_i t_j + \cdots\right)$$

となるから

$$f_n(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} n^{\frac{p}{2}} M(\hat{\boldsymbol{\theta}})^n e^{-n\hat{\boldsymbol{\theta}}'\boldsymbol{x}} |K_{ij}(\hat{\boldsymbol{\theta}})|^{-\frac{1}{2}} \left(1 + O\left(\frac{1}{n}\right)\right)$$

となる. 高次の展開も可能である (このことは既知である). 次に $\boldsymbol{y} \geq \boldsymbol{0}$ として

$$\begin{aligned} f_n\left(\boldsymbol{x} + \frac{\boldsymbol{y}}{n}\right) =& (2\pi)^{-\frac{p}{2}} n^{\frac{p}{2}} M(\hat{\boldsymbol{\theta}})^n e^{-n\hat{\boldsymbol{\theta}}'\boldsymbol{x} - \hat{\boldsymbol{\theta}}'\boldsymbol{y}} \\ & \times \int M\left(\hat{\boldsymbol{\theta}} + \frac{i\boldsymbol{t}}{n}\right)^n M(\hat{\boldsymbol{\theta}})^{-n} e^{-i\boldsymbol{t}'\boldsymbol{x} - i\boldsymbol{t}'\boldsymbol{y}/n} d\boldsymbol{t} \\ =& (2\pi)^{-\frac{p}{2}} n^{\frac{p}{2}} M(\hat{\boldsymbol{\theta}})^n e^{-n\hat{\boldsymbol{\theta}}'\boldsymbol{x}} |K_{ij}(\hat{\boldsymbol{\theta}})|^{-\frac{1}{2}} \\ & \times e^{-\hat{\boldsymbol{\theta}}'\boldsymbol{y}} \left(1 + \frac{1}{n}\boldsymbol{\kappa}'\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}'\boldsymbol{\Lambda}\boldsymbol{y} + O\left(\frac{1}{n^2}\right)\right) \end{aligned}$$

となるから, $\hat{\boldsymbol{\theta}} > \boldsymbol{0}$ ならば

$$\begin{aligned} P(\bar{\boldsymbol{X}}_n \geq \boldsymbol{x}) &= \int f_n\left(\boldsymbol{x} + \frac{\boldsymbol{y}}{n}\right) \frac{d\boldsymbol{y}}{n} \\ &= n^{-p} f_n(\boldsymbol{x}) \frac{1}{\prod \hat{\theta}_i} \left(1 + \frac{1}{n}\sum \frac{\kappa_i}{\hat{\theta}_i} + \frac{1}{n^2}\sum \frac{\Lambda_{ij}}{\hat{\theta}_i \hat{\theta}_j} + O\left(\frac{1}{n^2}\right)\right) \end{aligned}$$

となる. $\boldsymbol{X}_i$ の分布が正規分布であるとき, これは Mill's ratio の多次元への拡張を表す.

2) $\boldsymbol{X}_i$ の分布が格子点分布 $(X_i = 0, \pm 1, \cdots)$ のとき, 今度は $\boldsymbol{Y}_n = \sum_{i=1}^n \boldsymbol{X}_i$ とすれば

$$p(\boldsymbol{Y}_n = \boldsymbol{y}) = (2\pi)^{-p} M(\boldsymbol{\theta})^n e^{-\boldsymbol{\theta}'\boldsymbol{y}} \int_{-\pi\boldsymbol{1}}^{\pi\boldsymbol{1}} M(\boldsymbol{\theta} + i\boldsymbol{t})^n M(\boldsymbol{\theta})^{-n} e^{-i\boldsymbol{t}'\boldsymbol{y}} d\boldsymbol{t}.$$

また, $K_i(\hat{\boldsymbol{\theta}}) = y/n$ で $\hat{\boldsymbol{\theta}}$ を定義すれば

$$\begin{aligned} P(\boldsymbol{Y}_n = \boldsymbol{y}) &= (2\pi)^{-\frac{p}{2}} n^{-\frac{p}{2}} M(\hat{\boldsymbol{\theta}})^n e^{-\hat{\boldsymbol{\theta}}'\boldsymbol{y}} |K_{ij}(\hat{\boldsymbol{\theta}})|^{-\frac{1}{2}} \left(1 + O\left(\frac{1}{n}\right)\right) \\ P(\boldsymbol{Y}_n = \boldsymbol{y} + \boldsymbol{z}) &= P(\boldsymbol{Y}_n = \boldsymbol{y}) e^{-\hat{\boldsymbol{\theta}}'\boldsymbol{z}} \left(1 + O\left(\frac{1}{n}\right)\right). \end{aligned}$$

$\hat{\boldsymbol{\theta}} > \boldsymbol{0}$ ならば

$$P(\boldsymbol{Y}_n \geq \boldsymbol{y}) = \sum_{\boldsymbol{z} \geq \boldsymbol{0}} P(\boldsymbol{Y}_n = \boldsymbol{y} + \boldsymbol{z}) = P(\boldsymbol{Y}_n = \boldsymbol{y}) \sum \frac{1}{1 - e^{-\hat{\theta}_i}} \left(1 + O\left(\frac{1}{n}\right)\right).$$

Distributional and Related Problems on the Complex Matrix Spaces

香川大・工学部　　筑瀬靖子

1 Introduction

Let $\tilde{S}_m$ and $\tilde{R}_{m,p}$ be the spaces of $m \times m$ Hermitian matrices and of $m \times p$ complex rectangular matrices, respectively, that is,

$$\tilde{S}_m = \{S = S_1 + iS_2,\ with\ S_1\ m \times m\ symmetric\ and\ S_2\ m \times m\ skew - symmetric\}$$

and

$$\tilde{R}_{m,p} = \{Z = Z_1 + iZ_2,\ with\ Z_1\ and\ Z_2\ m \times p\ real\ rectangular\ matrices\}.$$

The statistics on $\tilde{S}_m$ and $\tilde{R}_{m,p}$ are of great use in Statistics, in particular, in time series analysis. This paper is concerned with some results on (i) complex matrix analysis, (ii) complex matrix-variate normal distributions defined on $\tilde{S}_m$ and $\tilde{R}_{m,p}$, and (iii) complex-Hermite polynomials associated with these complex matrix-variate normal distributions.

Some applications are presented to asymptotic distribution theory, defining certain exponential distributions on the complex Stiefel and Grassmann manifolds, in use of the theoretical results, especially Rodrigues formulae.

Some of the recent results on the complex normal and Wishart distributions are given by Andersen et al. [1]. Chikuse gives discussions of the real spaces [2, 4, 5] for the fundamental distribution theory and [3, 6] for the asymptotic distribution theory.

2 Complex Matrix-Variate Normal Distributions

2.1 Complex Normal $\tilde{\mathrm{N}}_{\mathrm{mm}}(0, \mathrm{I}_\mathrm{m})$ Distribution on $\tilde{\mathrm{S}}_\mathrm{m}$

The complex normal $\tilde{N}_{mm}(0, I_m)$ distribution defined on $\tilde{S}_m$ has the probability density

$$\tilde{\varphi}^{(m)}(S) = \pi^{-m^2/2} etr(-S^2) \tag{2.1}$$

with respect to the Lebesgue measure $(dS) = (dS_1)(dS_2)$, having the characteristic function $E\ etr(iTS) = etr(-\frac{1}{4}T^2)$ for $T \in \tilde{S}_m$. The general normal $\tilde{N}_{mm}(0, \Sigma)$ distribution is defined by $S = \Sigma^{1/2} V \Sigma^{1/2}$ with V being distributed as normal $\tilde{N}_{mm}(0, I_m)$, for $\Sigma \in \tilde{S}_m^+$ the space of $m \times m$ positive definite Hermitian matrices.

2.2 Complex Normal $\tilde{\mathrm{N}}_{\mathrm{m,p}}(0, \mathrm{I}_\mathrm{m} \otimes \mathrm{I}_\mathrm{p})$ Distribution on $\tilde{\mathrm{R}}_{\mathrm{m,p}}$

An $m \times p$ matrix $Z \in \tilde{R}_{m,p}$ is said to have the complex normal $\tilde{N}_{m,p}(0, I_m \otimes I_p)$ distribution if the probability density is given by $\varphi^{(m,p)}(Z) = \pi^{-mp} etr(-Z^*Z)$, with respect to the Lebesgue measure $(dZ) = (dZ_1)(dZ_2)$ (see [1]). The general normal $\tilde{N}_{m,p}(0, \Sigma_1 \otimes \Sigma_2)$ distribution is defined by $Z = \Sigma_1^{1/2} Y \Sigma_2^{1/2}$ with Y being distributed as normal $\tilde{N}_{m,p}(0, I_m \otimes I_p)$, for $\Sigma_1 \in \tilde{S}_m^+$ and $\Sigma_2 \in \tilde{S}_p^+$.

3 Associated Complex-Hermite Polynomials

3.1 Complex-Hermite Polynomials on $\tilde{S}_m$

We may define the complex-Hermite polynomials $\tilde{H}_\lambda^{(m)}(S)$ associated with the normal $\tilde{N}_{mm}(0, I_m)$ distribution, having the generating function

$$\sum_{l=0}^{\infty}\sum_{\lambda\vdash l}\frac{1}{\tilde{C}_\lambda(I_m)l!}\tilde{H}_\lambda^{(m)}(S)\tilde{C}_\lambda(T) = \int_{\tilde{O}(m)} etr(SH^*TH - \frac{1}{4}T^2)[dH],$$

where $\tilde{C}_\lambda(T)$ complex zonal polynomials in $T \in \tilde{S}_m$ with ordered partitions $\lambda \vdash l$ of an integer l, $\lambda = (l_1, \cdots, l_m), l_1 \geq \cdots \geq l_m \geq 0, \Sigma_{i=1}^m l_i = l$, and $[dH]$ the normalized invariant measure on the unitary group $\tilde{O}(m) = \{H; H^*H = I_m\}$. We obtain the Fourier transform, Rodrigues formulae [differential and integral versions (inverse Fourier transform)], $\tilde{H}_\lambda^{(m)}(S)\tilde{\varphi}^{(m)}(S) = \tilde{C}_\lambda(-\frac{1}{2}\partial S)\tilde{\varphi}^{(m)}(S)$, and series expansion for $\tilde{H}_\lambda^{(m)}(S)$. Use is made of the discussion of differential operators and Taylor expansions. The normal density (2.1) and the associated polynomials $\tilde{H}_\lambda^{(m)}(S)$ can be defined as certain limits of the complex-Wishart density and associated complex-Laguerre polynomials.

3.2 Complex-Hermite Polynomials on $\tilde{R}_{m,p}$

We can proceed a similar discussion of the complex-Hermite polynomials $\tilde{H}_\lambda^{(m,p)}(Z)$ associated with the normal $\tilde{N}_{m,p}(0, I_m \otimes I_p)$ distribution, including the close relationship with the complex-Laguerre polynomials.

References

[1] Andersen, H.H., Højbjerre, M., Sørensen, D., and Eriksen, P.S. (1995). *Linear and Graphical Models (for the Multivariate Complex Normal Distribution)*. Springer, New York

[2] Chikuse, Y. (1986). Multivariate Meixner classes of invariant distribuions.*Linear Algebra Appl.*, 82, 177-200.

[3] Chikuse, Y. (1991). Asymptotic expansions for distributions of the large sample matirx resultant and related statistics on the Stiefel manifold. *J. Multivariate Anal.*, 39, 270-283.

[4] Chikuse, Y. (1992). Properties of Hermite and Laguerre polynomials in matrix argument and their applications. *Linear Algebra Appl.*, 176, 237-260.

[5] Chikuse, Y. (1994). Generalized noncentral Hermite and Laguerre polynomials in multiple matrices. *Linear Algebra Appl.* 210, 209-226.

[6] Chikuse, Y., and Watson, G.S. (1995). Large sample asymptotic theory of tests for uniformity on the Grassmann manifold. *J. Multivariate Anal.* 54, 18-31.

Gassiat inequalities and some new results for singular models. (Abstract)

Didier Dacunha-Castelle Paris-Sud University

We give a global and simplify approach to test the order of a non-identifiable model using L.R.T or to identify this order using penalized likelihood. Then we these results to popular or new examples.

Let a family $(f_\theta)_{\theta\in\Gamma}$ of densities, Γ being the set of parameters in R^s and let Γ_0 the set of true and unknown values of θ and f_0 the true density. The singularity comes from the fact that Γ_0 is not a singleton and we are interested in situations as θ tends to Γ_0 and in the behavior of the likelihood or related functions.The more usual case is $\Gamma = \cup_{p\in P}\Gamma_p$ with a partial order on P and to test $p \prec q$ using LRT or to estimate p using penalized likelihood.

There are mainly to linked approaches. The first one, introduced by Dacunha-Castelle and Gassiat, uses a functional reparametrization using the set D of directionnal scores defined as $L^2(fdv)$ limits for θ tends to Γ_0 of $s_\theta = \frac{f_\theta - f_0/f_0}{\|f_\theta - f_0/f_0\|_{L^2(f_0)}}$,

D is included in the sphere of L^2, we call locally conic parametrization (LCP).The other approach is to find suitable expansions of the likelihood using specific distances as Hellinger or Pearson one,it is for instance the work of Liu and Shao. Using Gassiat powerful inequalities on the likelihood, and starting from the LCP, we get a nice expansion of the likelihood under natural hypothesis and make synthesis between the two approaches .

Gassiat inequalities. Let $s_g = \frac{g-f/f}{\|g-f/f\|_2}$ the norm being that of $L^2(fdv)$ and let ln(g) log-likelihood of the of the iid random variables $X_1..X_n$ when is their density. Let $(s_g)_- = -\min(0, s_g)$

Inequality 1 $\quad \sup_{g\in G:\ln(g)-\ln(f)\geq 0}\|g-f/f\|_2 \leq 2\sup_{g\in G}\frac{\sum_1^n s_g(X_i)}{\sum_1^n (s_g)_-^2(X_i)}$

Inequality 2

$$sup_{g\in G}(ln(g)-ln(f)) \leq \frac{1}{2}\sup_{g\in G}\frac{\left(\sum_1^n s_g(X_i)\right)^2}{\sum_1^n (s_g)_-^2(X_i)}$$

Let now G be a set of densities and $S = (s_g; g \in G)$. Let $H_{b,2}$ the entropy with bracketing of S with respect to the norm of $L^2(fv)$ and suppose $\int_0^1 \sqrt{H_{b,2}(u)}\,du < \infty$ so S is a Donsker class. Let D as previously defined by $d \in D$ iff it exists a sequence g_n such that $\|g_n - f_0/f_0\|_{L^2(f_0)} \to 0$ and $\|d - s_{g_n}\| \to 0$

Theorem (Gassiat) Under the previous hypothesis

$$sup_{g\in G}(ln(g)-ln(f)) = \frac{1}{2}\sup_D\left(\max(\frac{1}{\sqrt{n}}\sum^n d(X_i);0\right) + o_P(1)$$

From this result we obtain the classical limit theorem with for limit

$sup_D\ (\max(W(d);0)^2$ where W is a centered gaussian field with the scalar product in $L^2(f_0 v)$ as covariance.

Now we can applied this theorem in many situations, the main point being to generalize Gassiat inequalities and entropy estimates to more general situations than iid random variables.

Let us recall some examples studied in our group.

Order of finite Mixtures : Dacunha-Castelle and Gassiat

Mixtures with Markov regime : order estimation Gassiat and LRT Gassiat and Keribin

ARMA processes : order estimation and LRT:Dacunha-Castelle and Gassiat

A different situation is that of segmented regression models in Fedder sense.

Let $F = (f(\theta,t);\theta \in \Theta)$ a set of functions on [0, 1]and g a density of probability. Let $X_i = \sum_1^p f(\theta_j,t)1_{[t_{i-1},t_i]}(t) + \varepsilon_i$ be a segmented regression model with ε_i sequence of iid random variables with density g and $f(\theta_j,t) \in F$. The parameters are the order p the θ_j 's and the

break points t_j.

The model is clearly singular even, and it is and interesting case, if we suppose that $\sum_1^p f(\theta_j, t) 1_{[t_{i-1}, t_i]}(t)$ is a continuous function (note for instance in the case we want to test p=1 against p=2 the paper played by $t_1 = 0$ or $t_2 = 1$. For simple hypothesis on F we can apply the previous scheme of proof, after extension of the various ingredients to this case of independent variables with the same density after a suitable translation (depending on j).

References

Dacunha-Castelle, D and Gassiat, E. Testing in locally conic models. ESAIM prob. and stat., 1, 1997.

Dacunha-Castelle, D. and Gassiat, E. Testing the order of a model using locally conic parametrization : population mixtures and stationary ARMA processes. Annals of Stat., 27, 4 1178-1209, 1999.

Gassiat, E. Likelihood ratio inequalities with applications to various mixtures, preprint Orsay, to appear in Annals of statistics.

Leroux, B.G. and Puterman, M.L. Maximum-penalized likelihood estimation for independent and Markov dependent mixture models. Biometrics, 48, 545-558, 1992

Liu, X., Shao, Y. Asymptotics of the likelihood ratio under loss of identifiability with applications to finite mixture models. preprint, 2001

Likelihood Ratio of Unidentifiable Models and Multilayer Neural Networks

Kenji Fukumizu
The Institute of Statistical Mathematics
e-mail:fukumizu@ism.ac.jp

I discuss the asymptotic behavior of MLE under the condition that the true parameter is unidentifiable such as mixture models, ARMA, RRR, and neural networks. If the set of true parameters is of dimension larger than zero, the Fisher information matrix at a true parameter is singular, and the standard asymptotic normality is no longer satisfied.

Let $S = \{f(z;\theta) \mid \theta \in \Theta\}$ be a statistical model, and $Z_1, \ldots, Z_n$ i.i.d. sample from the true probability density $f_0 \in S$. I discuss on the asymptotic order of the likelihood ratio (LR) test statistics of MLE,

$$\sup_{\theta\in\Theta} L_n(\theta), \quad \text{where} \quad L_n(\theta) = \sum_{i=1}^n \log \frac{f(Z_i;\theta)}{f_0(Z_i)}, \tag{1}$$

as the sample-size n goes to infinity. This work focuses on divergence of LR in locally conic models (Dacunha-Castelle and Gassiat 1997), which formulate the unidentifiability of true parameters.

Locally conic models and divergence of LR

Let A_0 be a $(d-1)$-dimensional differentiable manifold, and Θ a submanifold in $A_0 \times \mathbb{R}_{\geq 0}$. The parameter $\theta \in \Theta$ is decomposed as $\theta = (\alpha, \beta)$ for $\alpha \in A_0$ and $\beta \in \mathbb{R}_{\geq 0}$. The statistical model S is called *locally conic* at f_0 if [1] Θ includes $\Theta_0 := A_0 \times \{0\}$, and the set of the parameters to give f_0 is Θ_0, [2] for each $\alpha \in A_0$, the set $\Theta(\alpha) := \{\beta \in \mathbb{R}_{\geq 0} \mid (\alpha, \beta) \in \Theta\}$ is a closed interval with open interior, and [3] for each α, $\left\| \frac{\partial}{\partial \beta} \log f(z;\alpha,0) \right\|_{L^2(f_0\mu)} = 1$. Intuitively, a locally conic model S is a union of one-dimensional submodels $S_\alpha = \{f(z;\alpha,\beta) \mid \beta \in \Theta(\alpha)\}$.

Under the assumptions of asymptotic normality for each S_α, the LR in the model S can be decomposed into (Dacunha-Castelle and Gassiat 1997)

$$\sup_{\theta\in\Theta} L_n(\theta) = \sup_{\alpha\in A_0} \left\{ \frac{1}{2} \left\{ \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n v_\alpha(Z_i) \right)_+ \right\}^2 + o_p(1) \right\}, \tag{2}$$

where $v_\alpha(z) = \frac{\partial}{\partial \beta} \log f(z;(\alpha,0))$ is a unit tangent vector along S_α. If we can find an arbitrary number of almost uncorrelated tangent vectors, the limiting

distributions of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} v_\alpha(Z_i)$ are almost independent Gaussian variables and the supremum is arbitrarily large. Generalizing this idea of Hartigan on a Gaussian mixture model (Hartigan 1985), we have the following useful sufficient condition of LR divergence;

Theorem 1. *Assume $S = \{f(z;(\alpha,\beta))\}$ is locally conic at $f_0 \in S$, and for each $\alpha \subset A_0$ the submodel $S_\alpha = \{f(z;\alpha,\beta) \mid \beta\}$ satisfies asymptotic normality. If there exists a sequence $\{v_n\}_{n=1}^{\infty}$ in the set of unit scores $\{\frac{\partial}{\partial\beta}\log f(z;(\alpha,0)) \mid \alpha \in A_0\}$ such that $v_n \to 0$ in probability, then, for arbitrary $M > 0$, we have*

$$\lim_{n\to\infty} \mathrm{Prob}\Big(\sup_{(\alpha,\beta)} L_n(\alpha,\beta) \leq M\Big) = 0. \tag{3}$$

Asymptotic order of LR in multilayer perceptrons

Another main result is the asymptotic order of LR for the multi-layer neural network model, which is defined by regression using the function family

$$\varphi(x;\theta) = \sum_{j=1}^{H} b_j \tanh(a_j x + c_j) + d, \tag{4}$$

where $x \in \mathbb{R}$ and $\theta = (a_1, b_1, c_1, \ldots, a_H, b_H, c_H, d) \in \mathbb{R}^{3H+1}$. I assume a law $Q = q(x)\mu_{\mathbb{R}}$ for input sample, and a conditional probability density function $r(y \mid u)$ for a noise model. The statistical model is defined by $f(x,y;\theta) = r(y|\varphi(x;\theta))q(x)$. A sample is given by the true density $r(y|\varphi_0(x))q(x)$ for the true function $\varphi_0(x)$. It is easy to see that the true parameter is unidentifiable if the true function is given by a network with a smaller number of hidden units than H. We can introduce a locally conic parameterization in this unidentifiability, and show divergence of LR as follows;

Theorem 2. *Assume that the model is the multilayer perceptron with H hidden units, and the true function is given by a network with K hidden units for $K < H$. Under some regularity conditions on the noise model $r(y|u)$, for arbitrary $M > 0$, we have*

$$\lim_{n\to\infty} \mathrm{Prob}\Big(\sup_{(\alpha,\beta)} L_n(\alpha,\beta) \leq M\Big) = 0. \tag{5}$$

If there are at least two redundant hidden units to realize the true function, we can derive a lower bound of the order of LR.

Theorem 3. *Assume that the model is the multilayer perceptrons with H hidden units, and the true function is given by a network with K hidden*

units for $K \leq H-2$. Under some regularity conditions on the noise model $r(y|u)$, there exists $\delta > 0$ such that

$$\liminf_{n\to\infty} \mathrm{Prob}(\sup_\theta L_n(\theta) \geq \delta \log n) \quad > \quad 0. \tag{6}$$

If the noise model is Gaussian, we have $\log n$ upper bound for a wide class of regressors.

Theorem 4. *Assume that the VC dimension of a function class $\mathcal{F}$ is finite, and the true function $\varphi_0 \in \mathcal{F}$ is bounded. Then, for the statistical model $\{f(x,y;\varphi) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(y-\varphi(x))^2\} q(x) \mid \varphi \in \mathcal{F}\}$, we have*

$$\sup_{\varphi\in\mathcal{F}} \sum_{i-1}^{n} \log \frac{f(X_i,Y_i;\varphi)}{f(X_i,Y_i;\varphi_0)} = O_p(\log n). \tag{7}$$

From the above two theorems, we know that LR of the multilayer perceptron model is of exactly order $\log n$, if the model has at least two redundant hidden units to realize the true function and the noise is Gaussian.

Most of the results presented here are completely shown in Fukumizu (2001) (http://www.ism.ac.jp/~fukumizu/papers/memo780.pdf).

References

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

Fukumizu, K. (2001). Likelihood ratio of unidentitable models and multilayer neural networks. Research memorandum, Institute of Statistical Mathematics.

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, pp. 807–810.

特異点解消および確率的複雑さの法則収束

東京工業大学　渡辺澄夫

1　問題

密度関数 $q(x)$ を持つ確率変数 X からの n 個の独立なサンプルを $X^n = (X_1, X_2, \cdots, X_n)$ と書く。d 次元多様体の部分集合 W をパラメータ空間として持つパラメトリックモデル $p(x|w)$ $(w \in W)$ による統計的推測の問題を考える。特に $w \mapsto p(\cdot|w)$ が1対1でないとき、特定不能な統計モデルという。カルバック情報量を

$$H(w) = \int q(x) \log \frac{q(x)}{p(x|w)} dx$$

とする。$H(w) = 0$ を満たすパラメータ（真のパラメータ）の集合が次元を持つ広がりとなるとき、統計的推測の精度がどのようになるかは未解決の問題である。この問題は、混合正規分布や神経回路網などの、外界からは隠れて見えない部分を持つ確率モデルでは必ず現れる問題であり、特に、知能情報科学で多用されている複雑な構造を持つ推論モデルにおいて避けては通れない問題である。この報告では、ベイズ法の推測精度の漸近論について一般的解答を与え、その背後にある数理的な構造を考える。

2　ベイズ推測

パラメータ空間に事前分布 $\varphi(w)$ があるとする。これから作られる事後分布を $p(w|X^n)$ とし、ベイズ予測分布を $p(x|X^n) = \int p(x|w)p(w|X^n)dw$ とする。ベイズ推測においては、確率的複雑さ（タイプII対数尤度の符号反転）

$$F(X^n) = -\log \int \prod_{i=1}^{n} p(X_i|w)\varphi(w)dw$$

が次の3点で重要な役割を果たす。(1) ベイズ予測誤差は、確率的複雑さの増分に等しい。すなわち、$-\log p(X_{n+1}|X^n) = F(X^{n+1}) - F(X^n)$ が成り立つ。(2) ベイズモデル選択は、確率的複雑さを最小にすることにより行われる。(3) ハイパーパラメータは確率的複雑さの最小化により最適化される。そこで確率的複雑さの漸近的なふるまいを解明しよう。

3　結果

本報告の主要定理は次の通りである。

定理 1 カルバック情報量 $H(w)$ が w の解析関数であるならば（より厳密には、自然な仮定を幾つか付加することにより）、確率的複雑さは $n \to \infty$ で次のように漸近展開できる。

$$F(X^n) = S(X^n) + \lambda \log n - (m-1) \log \log n + R(X^n) \qquad (1)$$

ここで $S(X^n) = -\sum_{i=1}^{n} \log q(X_i)$ は経験エントロピーであり、λ と m はゼータ関数

$$J(z) = \int H(w)^z \varphi(w)$$

の最も絶対値の小さい極とその位数である。ここで $J(z)$ は全複素平面上で定義できる有理型関数で、その極はすべて負の有理数である。また $R(X^n)$ は、真のパラメータ集合の上の経験過程の極限として得られるガウス過程の積分で表される確率変数に法則収束する。

このことの証明は、特異点解消定理 (広中によって証明された。論文 (Atyah,1970) の解説が簡潔でわかりやすい) に基づいて $H(w) = 0$ の特異点を解消し、ゼータ関数を通して確率的複雑さを経験過程で表すことによって行われる。λ と m は、$H(w) = 0$ の特異点解消を得ることで求められる。特異点の解消は必ず存在するが、その発見は困難であることが少なくない。しかしながら $J(z)$ の一般の極 $-\lambda'$ は、ブローアップによって求めることができ、$-\lambda$ が最も絶対値の小さい極であることから、$\lambda \leq \lambda'$ が成り立つので、上限を得ることは容易である。一般に $1 \leq m \leq d$ が成り立ち、もしも、事前分布が真のパラメータ上で 0 でなければ、$0 < \lambda \leq d/2$ が成り立つ。Jeffreys 事前分布は真のパラメータ上で 0 になり、一般に $\lambda \leq d/2$ が成り立つ。

4 考察

以上の問題に関連して、今後の課題を述べる。(1) 特異点解消定理は統計学では応用されたことがない。本報告で述べた定理の真価を、より多くの方にご理解いただくために代数幾何を用いない証明が望まれる。(2) 実問題への応用では、真の密度関数は統計モデルに完全には含まれていない。その場合に生ずる現象を解明する。(3) λ と m については、幾何学的な解釈を与えることができるが、$R(X^n)$ は未だに不明である。これに幾何学的な解釈を与える。(4) Jeffreys の事前分布は、特異モデルのモデル選択において有用であると思われるが、$\lambda = d/2$ とならない統計モデルも存在する。Jeffreys の事前分布よりも、普遍的な事前分布は存在するのだろうか。(5) 本論文では確率的複雑さの漸近論を与えたが、確率的複雑さを具体的に計算するアルゴリズムを与えたわけではない。パラメータが特異点を持つ場合において有効となる計算手法を確立する。

5 結論

特異点を持つ学習モデルのベイズ推測における漸近論を初めて確立した。ゼータ関数によって、代数幾何と統計学とが数理的に結ばれ、神経回路網などの大規模で高度に複雑な確率モデルの解析を行うためのひとつの基盤が構成された。

参考文献

Atiyah, M. F. (1970). Resolution of Singularities and Division of Distributions. *Communications of Pure and Applied Mathematics* , 13, 145-150.

Watanabe, S. (2001). Algebraic analysis for non-identifiable learning machines. *Neural Computation*, Vol.13,No.4, 899-933.