ULAR LEARNING THEORY: CONNECTING ALGEBRAIC GEOMETRY AND MODEL SELECTION IN STAT

organized by
Russell Steele, Bernd Sturmfels, and Sumio Watanabe

Workshop Summary

1 Introduction

This article reports the workshop, "Singular learning theory: connecting algebraic geometry and model selection in statistics," held at American Institute of Mathematics, Palo Alto, California, in December 12 to December 16, 2011. In this workshop, 29 researchers of mathematics, statistics, and computer science, studied the following issues: (i) mathematical foundation for statistical model selection; (ii)Problems and results in algebraic statistics; (iii) statistical methodology for model selection; and (iv) statistical models in computer science.

2 Background and Main Results

Many statistical models and learning machines have hierarchical structures, hidden variables, and grammatical information processes and are being used in vast number of scientific areas, e.g. artificial intelligence, information systems, and bioinformatics. Normal mixtures, binomial mixtures, artificial neural networks, hidden Markov models, reduced rank regressions, and Bayes networks are amongst the most examples of these models utilized in practice.

Although such statistical models can have a huge impact within these other disciplines, no sound statistical foundation to evaluate the models has yet been established, because they do not satisfy a common regularity condition necessary for modern statistical inference. In particular, if the likelihood function can be approximated by a multi-dimensional normal distribution, then the accuracy of statistical estimation can be established by using properties of the normal distribution. For example, the maximum likelihood estimator will be asymptotically normally distributed and the Bayes posterior distribution can also be approximated by the normal distribution.

However, many of the statistical models used in modern information science do not satisfy the basic regularity conditions, such that Fisher's asymptotic normal theory does not hold. As a result, model selection criteria such as AIC, BIC, and DIC have no theoretical support for this class of models. None of them correspond to the generalization error or the marginal likelihood, the two most popular objective criteria for model selection. The main reason of this difficulty is that the correspondence of the parameter and the probability distribution is not one-to-one and that the optimal set of parameters consists of an algebraic or an analytic set with singularities. Such statistical models are called *singular*. If a statistical model is devised to extract some hidden structure from random samples, then, in general, it is not regular but singular.

The main purpose of this workshop was to create the mathematical theory and methods to treat singular statistical problems using commutative ring theory, algebraic geometry, and algebraic analysis. The fundamental ideas from this *singular learning theory* are summarized in the four points below:

(1) Algebraic geometry and algebraic analysis give the essential solution of singular statistical problems.
(2) The asymptotic form of the log Bayes marginal likelihood is given by the real log canonical threshold.
(3) The difference between the generalization error and the training error is given by the singular fluctuation.
(4) Statistical model selection criteria, AIC, BIC, and DIC can be generalized so that they can be applicable to singular statistical models.

This workshop clarified that modern mathematics can be an extremely powerful tool for resolving statistically difficult problems and that modern statistics can be a fertile area for the generation of new mathematical theories and concepts.

3 Mathematical foundation

Let W be a compact subset in $\mathbb{R}^d$. A statistical model $p(x|w)$ is defined as the probability density function of $x \in \mathbb{R}^N$ for a given parameter $w \in W$. Assume that random variables $X_1, X_2, ..., X_n$ are independently subject to the same probability density function $q(x)$. In singular statistical theory, we mainly study the case that the set of true parameters

$$W_0 = \{w \in W \ ; \ q(x) = p(x|w)\}$$

is an algebraic or an analytic set with singularities in W. The minus log marginal likelihood or the free energy is a real-valued random variable defined by

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

where $\varphi(w)$ is a probability density function on W called a prior distribution. This random variable is one of the most important observable used in Bayes statistics. Professor I. J. Good proposed in 1960 that model selection and hyperparameter optimization can be done by minimization of F. Professor G. Schwarz showed in 1978 that it is determined by the half of the dimension of the parameter space when regularity condition is satisfied,

$$F = nS_n + \frac{d}{2}\log n + O_p(1),$$

where S_n is the empirical entropy of $q(x)$ and $O_p(1)$ is a random variable which converges to a random variable in law. However, its correct form for singular models has been left unknown.

In this workshop, we generalized this result onto the general cases that contain singular cases,

$$F = nS_n + \lambda \log n - (m-1)\log\log n + O_p(1), \tag{1}$$

where λ and m are a positive rational number and a natural number which are respectively determined by the zeta function

$$\zeta(z) = \int K(w)^z \varphi(w) dw,$$

using Kullback-Leibler distance $K(w)$ between $q(x)$ and $p(x|w)$. The zeta function is a holomorphic function in the region $\mathrm{Re}(z) > 0$, which is analytically continued to the unique meromorphic function on the entire complex plane. Its poles are all real, negative, and rational numbers. If its largest pole is $(-\lambda)$, then λ is the real log canonical threshold (RLCT). The order of the pole $(-\lambda)$ is denoted by m. In mathematics, RLCT is an important and well-known birational invariant. It is defined for two algebraic varieties W_0 and W and Professor Mustata showed that RLCT shows the relative properties of their singularities of the pair (W, W_0). Therefore the Bayes marginal is mathematically determined by the relative properties of the parameter set and the set of true parameters.

In algebraic analysis, the above zeta function has the mathematical relation to Bernstein-Sato polynomial or a b-function. There exists a set $(P_z, b(z))$ and $b(z)$ where P_z is a differential operator of w and $b(z)$ is a polynomial which satisfies

$$P_z K(w)^{z+1} = b(z) K(w)^z.$$

The minimal order function $b(z)$ which satisfies this relation is called Bernstein-Sato polynomial if the coefficient of the largest order term is one. For a given polynomial $K(w)$ there exists an algebraic algorithm to find $b(z)$.

In regular statistical models, RLCT is equal to the half of the number of parameters and $m = 1$, hence our result is a generalized and nontrivial one. It should be emphasized that equation (1) could not be found without algebraic geometry and algebraic analysis. In statistics and machine learning, the concrete values of RLCT are very important. For an arbitrary analytic function $K(w)$, there exist a parametrization $g : U \to W$ such that

$$K(g(u)) = u_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d},$$

where $k_1, k_2, ..., k_d$ are nonnegative integers, U is a subset of some $d-$dimensional manifold, and g is a proper analytic map. This theorem claims that an arbitrary analytic set can is an image of a normal crossing analytic set. This theorem is referred to as Hironaka's resolution theorem. Note that the Jacobian determinant of $w = g(u)$ is also given by the normal crossing function,

$$|g'(u)| = b(u)|u_1^{h_1} u_2^{h_2} \cdots u_d^{h_d}|,$$

where $b(u) > 0$ and $h_1, h_2, ..., h_d$ are nonnegative integers. If we can find such map $w = g(u)$, then the RLCT is given by

$$\min_{1 \le j \le d} \frac{h_j + 1}{2k_j},$$

where we put $1/k_j = \infty$ if $k_j = 0$. Therefore, if we can find the resolution of singularities, then we will obtain the RLCT. If we can obtain the RLCT, we can use it for approximation of the stochastic complexity. We can then use the approximation of the stochastic complexity for statistical model selection.

4 Summary of workshop presentations

There were 10 presentations given during the 5 mornings of the workshop. The presentations ranged from having a very mathematical focus to a very statistical focus, with most talks falling somewhere in between. Dr. Shaowei Lin's opening seminar on the Monday morning introduced singular learning theory to the mathematicians and the statisticians in the audience. In particular, Dr. Lin clarified the relation of RCLT and algebraic statistics. For example, Dr. Lin showed that if

$$\langle f_1, f_2, ...f_k \rangle = \langle g_1, g_2, ..., g_l \rangle$$

then two functions

$$K_1(w) = f_1(x)^2 + \cdots + f_k(x)^2$$
$$K_2(w) = g_1(x)^2 + \cdots + g_l(x)^2$$

have the same RLCT. By using this property and Hironaka's resolution theorem in algebraic geometry, we can find the concrete value of RLCT.

The second speaker on the first day, Dr. Mathias Drton, focused on a particular model selection problem in statistics, namely choosing the rank in a reduced rank regression model. His talk presented a novel approach for estimating this rank using singular learning theory. This problem was taken up by one of the working groups and further developed later in the day.

On the second day, Dr. Anton Leykin gave an overview of methods used in computational algebraic geometry for resolution of singularities, computations that are essential to calculating learning coefficients in singular models. During his talk, Dr. Leykin showed that the difference of RCLT and the minimum root of b-function is an integer. By using this result, he made a new hybrid method to find RLCT using combining D-module calculation with numerical integration. Dr. Sumio Watanabe gave the second Tuesday talk, providing an overview of the most recent results in singular learning theory and highlighted directions for future research.

Dr. Franz Kiraly and Dr. Helene Massam both presented new problems in singular learning theory on Wednesday. Dr. Kiraly's talk presented an approximate algebraic approach for estimation when the data under consideration are generated as random ideals from a variety. Dr. Massam presented a problem from log-linear models which requires estimation of the limit of an integral as a set of hyperparameters tends towards zero, which is a problem of a very different nature than the current asymptotics for integrals that depend on increasing sample size.

Dr. Martyn Plummer's talk on the Thursday morning presented more statistically oriented model selection criteria that are commonly used by Bayesian statisticians and discussed their relationship to singular learning criteria. During the second Thursday talk, Dr. Miki Aoygi walked the group through the basic principles behind deriving learning coefficients by hand for certain types of problems. Dr. Aoyagi showed how to find desingularization map using recursive blow-ups and showed that the concrete values of RLCT were discovered in several statistical models. For example, RLCT in reduced rank regression corresponds to

the largest pole of the zeta function

$$\zeta(z) = \int \|BA - C\|^{2z} dA dB,$$

where A, B, C are matrices and rank(C) is smaller than those of A and B. It was clarified that RCLT is obtained for arbitrary rank of C and that the result coincides with statistically numerical analysis. Also it was pointed out that, for more general statistical models such as normal mixtures and artificial neural networks, desingularization of Vandermonde matrix type singularities is necessary.

On the final day, Dr. Piotr Zwiernik presented his recently published results on the use of the Newton polytope approach for deriving learning coefficients in the case of Markov models. In particular, Dr. Zwiernik clarified that RLCT is necessary in model election of Bayesian network and created an algorithm to find RCLT for arbitrary Bayesian network. The final talk of the workshop was given by Dr. Alexander Schliep, who discussed where in bioinformatics singular learning theory might be useful.

As can be seen from these short descriptions, the talks balanced reviews and demonstrations of existing results with some new results and characterization of open problems. The presentations were used to generate ideas for the afternoon sessions.

5 Summary of afternoon group sessions

There were three sets of groups established. The first set of groups worked in the afternoons on Monday and Tuesday the second set of groups worked in the afternoons on Tuesday and Wednesday, with the final set of groups gathering on Thursday and Friday.

Three groups in the first round focused on the computation of integrals. One group focused on exact calculations, the second group focused on asymptotic approximation, and the third group focused on Monte Carlo approximation. Two other groups were also formed. One group discussed issues regarding knowledge transfer – how statisticians can be introduced to the algebraic geometric ideas and how algebraic geometers can be made aware of the challenging mathematical problems that arise in statistical modelling. The final group explored different "favorite" models in order to try to generate ideas for later groups and for future directions. During final reports of these groups, it was clear that there is much work to do with both knowledge transfer and in identifying useful statistical models that would be easily amenable to the use of singular learning theory. The Monte Carlo group made rapid process and generated computer code to estimate the Bayesian integrated likelihood for the reduced rank regression problem for a few special cases.

The second round of groups were focused on four topics: discrete hidden variable models, factor analysis models, hybrid computational methods, and understanding ν (a particular parameter that appears in singular learning models, but does not have a easily accessible geometric interpretation). The discrete hidden variable models group worked through some simple, but very illustrative examples of how Newton polytope methods can be used to derive

learning coefficients in this class of of models. The hybrid computational methods group implemented Dr. Leykin's Monte Carlo "blowing up" approach for a continuous mixture model and derived what is believed to be a new learning coefficient for this class of models. The factor analysis group made very good progress towards obtaining learning coefficients for lower dimensional models. The understanding ν group came to develop an idea of ν as representing the "thickness" of a variety, which would be connected to statistical ideas of variability or instability in the estimation of the statistical models.

The final round of groups focused on four topics: the connection between Bayesian estimates of model complexity (p_D) and the singular learning coefficient λ and the constant ν; the factor analysis group continued their derivations of learning coefficients for lower dimensional models; the discrete hidden variables model group then directed their focus to the loglinear model integrals discussed in Dr. Massam's talk; and the final group discussed Dr. Aoyagi's derivation of singularities by hand in the case of reduced rank regression and neural network models and attempted to derive the same learning coefficients using computational methods. All groups made significant progress by the final afternoon. The p_D group managed to connect ideas between three different versions of the Deviance Information Criterion and the WAIC from singular learning theory and establish conditions under which these criteria will yield the same selected (and true) models and when they will differ and how. For special cases, the singularities group was able to generate via a computational approach the same learning coefficients that were derived by hand in Dr. Aoyagi's work.

6 Summary of the problem session

On Thursday afternoon, the workshop held its problem session for proposing open problems and discussing where the field would go next. We identified four key open problems in singular learning, some of which have associated sub-problems, as well as some other potential future directions.

1. What are the properties of singular learning criteria if the true distribution is outside the families of models considered?

Singular learning theory currently assumes that the true distribution is inside the family of models under consideration. Therefore, one of the most important open problems in singular learning theory is to extend or understand the theory when this assumption does not apply or when it is only approximately true. Some examples of this include the following problems:

(1) Let us assume that the true probability distribution q of a data generating process is outside the family of models $\mathcal{M}$ that are being considered. Under what conditions does the posterior distribution converge to a distribution with smallest KL divergence to the true distribution q?
(2) What are the asymptotics of model selection via the stochastic complexity in the case where the true distribution is not in any model under consideration? This is only partially worked out for regular statistical models; the question remains open for singular models.

(3) One concrete example was suggested as a potential direction. Suppose the data are generated from a log normal distribution, and assume two potential families of distributions: a) a normal distribution and b) Gamma distribution. Given the simplicity of the mathematics for these three families of distributions (and potential simiarities), we may be able to understand the behavior of asymptotics in this particular case.
(4) **Y. Zhang:** As another example, let us assume the true model is a non-linear regression model. $Y = exp(X) + \epsilon$ where $\epsilon \sim N(\mu, \sigma^2)$. Let us assume that we try to fit a linear regression model to the data, i.e. $y = bx + \epsilon$. What is the behavior of the stochastic conmpleity in this case?
(5) **S. Watanabe:** When the true distribution is outside the model, when the model is fixed and the true distribution is fixed, then the distance between them is fixed and the distance is order 1. What is more interesting is to study the cases when the distance between them is order is $1/n$ where n is the number of training samples. In the case of the suggested particular example, we need to study (a) and (b) when the KL distance between true distribution (q) and the statistical model is prop to $1/n$. In this case variance and bias are almost equal. We need to compare statistical bias with functional approximation bias and there are no standard solutions.

2. Computing the learning coefficients

Approximation of the stochastic complexity using the methods of singular learning requires the computation of the learning coefficients. Much of the workshop was devoted to exploring different approaches (exact, analytic approximation, and Monte Carlo methods) to this problem for basic singular models. Some of the ideas that were suggested for open problems in this area are below:

(1) **M. Drton and S. Sullivant:** One useful exercise would be to find the learning coefficient λ for the factor analysis model and be able to describe the singularities of the factor analysis fibers. The advantage of working with the factor analysis model is its wide popularity in statistics and the ability to generalize it to more complex, but related latent variable models.
(2) **S. Watanabe** Another problem would be to determine RLCT by a mixed symbolic and numeric approach. i.e. hybrid approach.
(3) **A. Leykin:** One way to address the learning coefficient problem would be if we could, a prior, determine the denominator of the RLCT.
(4) **F. Kiraly:** An additional problem that could be addressed that would be somewhat different than the models currently addressed in the literature would be to find the learning coefficients for the ideal regression problem.

3. Mixture of Normal Distributions

Mixtures of normal distributions have been explored to great extent in the published research in singular learning. They are very commonly used in many areas of statistics, so this class of problems continued to generate interest during the course of the workshop. Some additional problems in mixture models were suggested during the course of the problem session.

(1) **S. Ray / B. Sturmfels:** A very interesting question was proposed by an invited participant (S. Ray) who was unable to attend: what is the maximum number of

modes for a mixture of k normal distributions in dimension d? It is known for $k = 2$, and $k = d = 3$. It would be a very interesting and different application of singular learning to answer this question. B. Sturmfels' conjecture is that the maximum number of modes is $\binom{d+k+1}{k-1}$. Some progress was made on this during the working group, however the problem remains unsolved.

4. Applications of using learning coefficients for model selection

The role of the learning coefficient in singular learning is very important with regards to model selection via the stochastic complexity. Therefore, beyond simply computing the learning coefficient, it is critical to understand what the statistical interpretation of the learning coefficient would be in practice and how it relates to other measures of model complexity in statistics.

(1) **M. Drton** There was much discussion of the two-step procedure that was proposed in Dr. Drton's talk. One of the most immediate open problems is to finalize the two step procedure for model selection in Reduced Rank regression and examine its behaviour in simulations.
(2) **V. Karwa:** After the reduced rank regression two-step procedure is finished, there could be a potential extension for the problem of choosing the number of components in a D dimensional Gaussian mixture model.
(3) **R.Steele:** What is the relationship between the penalty term that Dr. Plummer presented in his talk, p_d^* (effective degrees of freedom) and ν or λ? A working group examined this in the final two days, but it was not formally established.
(4) **S. Watanabe:** In many situations, there is only a known bound for the learning coefficient λ. In these situations, can we use Dr. Drton's two step model selection method?

5. Other Problems

In addition to the four areas discussed above, these additional open problems and future directions were suggested by participants.

(1) **J. Morton:** Can on edetermine λ as a function for homogeneously parametrized undirected graphical models with at most 4 (possibly hidden) nodes?
(2) **B. Sturmfels:** What is the nature of the problem if the true distribution is in the boundary of the parameter space?
(3) **L. Pericchi:** Can we understand the asymptotics of the stochastic complexity in the case of dependent data (more generally not identically distributed)?
(4) **B. Sturmfels:** With regards to computing the Bayes factor for toric models, one could try to calculate $J_c(M)$ for the binary m-cycle exactly. An additional problems would be to compute the $I_(M, \alpha)$ which is the marginal likelihood or the stochastic complexity.
(5) **R. Steele:** Under what conditions can we obtain results for the asymptotic stochastic complexity when D increases with n? As an example, let $X_{ij} \sim N(\mu_j, \sigma^2)$ where $i = 1, \ldots n_j$. Let $\mu_j \sim N(\mu, \tau^2)$ where $j = 1, \ldots, k$. Let n_j and k go to infinity. Here $D = k+2$, assuming that σ^2 is constant, and the sample size is $\sum_i n_i$. Now let n_j and k be large. What is the behaviour of the stochastic complexity in these situations? Can the results of singular learning be modified to still give useful statistical inference?

(6) **S. Watanabe:** There should be other approaches to study Dr. Kiraly's proposed problem for estimating random ideals statistically. It would be interesting to explore other methods to estimate ideal from random noisy samples from the ideal.
(7) **S. Watanabe:** It is very important to clarify the meaning of ν. What are the algebro-geometric and statistical meanings of the singular fluctuation ν? Does it have any relation to any birational invariant?

Acknowledgment. We would like to thank the members of this workshop, Miki Aoyagi, Arnon Boneh, Andrew Critch, Mathias Drton, Luis Garcia-Puente, Helene Gehrmann, Elizabeth Gross, Vishesh Karwa, Franz Kiraly, Gerard Letac, Anton Leykin, Shaowei Lin, Helene Massam, Guido Montufar, Jason Morton, Luis Pericchi, Sonja Petrovic, Martyn Plummer, Jose Rodriguez, Alexander Schliep, Aleksandra Slavkovic, Seth Sullivant, Caroline Uhler, Jing Xi, Yongli Zhang, and Piotr Zwiernik.

Bibliography

[Akaike] Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, 19, 16-723.

[Atiyah] Atiyah,M.F. (1970) Resolution of singularities and division of distributions. *Communications of Pure and Applied Mathematics*, 13, 145-150. 1

[Bernstein] Bernstein,I.N. (1972) The analytic continuation of generalized functions with respect to a parameter. *Functional Analysis and Applications*, 6 pp.26-40.

[Drton] Drton,M., Sturmfels,B., and Sullivant,S. (2009) *Lecures on Algebraic Statistics.* Birkhäuser, Berlin.

[Gelfand] Gelfand,I.M. and Shilov,G.E. (1964) *Generalized Functions.* Academic Press, San Diego.

[Gelman] Gelman,A., Carlin,J.B., Stern,H.S., Rubin.D.B.(2004) *Bayesian data analysis.* Chapman & Hall CRC, Boca Raton.

[Hironaka] Hironaka,H. (1964) Resolution of singularities of an algebraic variety over a field of characteristic zero. *Annals of Mathematics,* Vol.79, pp.109-326.

[Kashiwara] Kashiwara,K. (1976) B-functions and holonomic systems. *Inventiones Mathematicae,* 38, 33-53.

[Mustata] Mustata,M. (2002) Singularities of pairs via jet schemes, J. Amer. Math. Soc. 15, 599-615.

[DIC] Spiegelhalter,D.J., Best,N.G., Carlin,B.P., and Linde.A. (2002) Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society, Series B*, 64, 4, 583-639.

[Wvan] Van der Vaart,A.W. and Wellner,J.A. (1996) *Weak Convergence and Empirical Processes.* Springer.

[Cambridge2009] Watanabe,S.(2009) Algebraic geometry and statistical learning theory, Cambridge University Press, Cambridge, UK.