Singular learning theory: algebraic geometry and model selection

The American Institute of Mathematics

The following compilation of participant contributions is only intended as a lead-in to the AIM workshop "Singular learning theory: algebraic geometry and model selection." This material is not for public distribution.

Corrections and new material are welcomed and can be sent to `workshops@aimath.org`

Version: Mon Nov 28 18:20:14 2011

Table of Contents

Chapter A: Participant Contributions

A.1 Aoyagi, Miki

We consider a real log canonical threshold of singularities in learning theory. Such a threshold corresponds to a learning coefficient of generalization error in Bayesian estimation, which serves to measure the learning efficiencies in hierarchical learning models. We overview the learning coefficients obtained by us and also give a new results for Vandermonde matrix type singularity.

A.2 Boneh, Arnon

I have strong interst in black box and pink box input-output models which can be represented by the Volterra convolution series. Of special interest to me are the slightly nonlinear models of conservative systems where the distinction should be made between time varying and stationary models in the presence of large amounts of measured noisy data. This leads to large size non convex quadratically constrained quadratic programming problems with a high rate of built in redundancy. The model selection problem in such problems and the learning of model properties online is what connects me to the workshop in which I would like to participate.

A.3 Drton, Mathias

Watanabe's work provides a general understanding of the asymptotic growth behavior of marginal likelihood integrals in Bayesian approaches to model selection. I hope that during the workshop we can discuss how knowledge about the learning coefficients = (real) log-canonical thresholds in Watanabe's theory could be exploited in model selection. To my knowledge, no practical proposal for a statistical method currently exists. I will try to briefly explain the issues involved.

Suppose we are given independent and identically distributed observations $X_1, \ldots, X_n$ and competing models $\mathcal{M}_1, \ldots, \mathcal{M}_p$ for the distribution of these observations. Suppose that model $\mathcal{M}_i$ has likelihood function $L_i(\theta \,|\, X_1, \ldots, X_n)$ and that, in a Bayesian approach, the parameter θ in this model has prior distribution $q_i(\theta)$. If p_i denotes the prior probability, then the posterior probability of model $\mathcal{M}_i$ is proportional to

$$p_i \times \int L_i(\theta \,|\, X_1, \ldots, X_n) q_i(\theta)\, d\theta.$$

The integral just written is known as the marginal likelihood of model $\mathcal{M}_i$.

In his book, Watanabe derives the asymptotic growth behavior of marginal likelihood integrals under mild regularity conditions on the model $\mathcal{M}_i$ and technical compactness assumptions on the support of the prior q_i. In the asymptotic study, it is assumed that the observations $X_1, X_2, \ldots, X_n$ are identically distributed according to a distribution in $\mathcal{M}_i$. Write θ_0 for the parameter indexing this *true* distribution. The marginal likelihood integral

$$M_i(n) := \int L_i(\theta \,|\, X_1, \ldots, X_n) q_i(\theta)\, d\theta$$

is then a random variable and Watanabe's study treats the asymptotic behavior of the sequence $(M_i(n))_{n=1}^{\infty}$. He proves results that state that, as $n \to \infty$,

$$\log M_i(n) = \log L_i(\hat{\theta}_n \,|\, X_1, \ldots, X_n) - \lambda(\theta_0) \log n + [m(\theta_0) - 1] \log\log n + R_n,$$

where $\hat{\theta}_n$ is the maximum likelihood estimator, the constant $\lambda(\theta_0)$ is the so-called learning coefficient, $m(\theta_0)$ is its order, and R_n is a sequence of random variables that converges in distribution to a stochastic remainder term.

In the classical case of smooth models, the learning coefficient $\lambda(\theta_0)$ does not depend on the true parameter θ_0 and equals $1/2$ times the dimension of the model. Moreover, the multiplicity $m(\theta_0)$ is always equal to one. Therefore, neglecting the (probabilistically) bounded remainder term we may select a model by maximizing the score

$$p_i \times \left(\log L_i(\hat{\theta}_n \,|\, X_1, \ldots, X_n) - \frac{\dim(\mathcal{M}_i)}{2} \log n \right).$$

Often the prior probabilities p_i are uniform in which we simply maximize the so-called Bayesian information criterion (BIC)

$$\log L_i(\hat{\theta}_n \,|\, X_1, \ldots, X_n) - \frac{\dim(\mathcal{M}_i)}{2} \log n. \tag{1}$$

(In the smooth case, it is also known that the Hessian of the log-likelihood function at the maximum likelihood estimator can be used to estimate the remainder term R_n, which leads to an actual approximation to the marginal likelihood.)

However, Watanabe's result also covers singular models. In singular models the learning coefficient $\lambda(\theta_0)$ and its order $m(\theta_0)$ typically depend on the *unknown* parameter θ_0. Hence, it is not clear how learning coefficients can be used to define a model score to optimize. It is tempting to use the learning coefficient at the maximum likelihood estimator, that is, $\lambda(\hat{\theta}_n)$, but in some singular models the maximum likelihood estimator is a smooth point with probability one, which means that we are back at considering the classical BIC as defined in (1).

One model that is perfectly suited for experimentation is the reduced rank regression model treated in the paper of Aoyagi and Watanabe (2005):

`http://dx.doi.org.proxy.uchicago.edu/10.1016/j.neunet.2005.03.014`

This paper fully solves the problem of determining all learning coefficients together with their orders. It is an instance where maximum likelihood estimators are smooth points with probability one. I would like to ask the participants for their thoughts on how to use the mathematics from Aoyagi and Watanabe (2005) to create an improved information criterion for model selection in reduced rank regression.

A.4 Garcia-Puente, Luis

For the last several years I have been intersested in modifications of the Bayesian Information Criterion and the Akaike Information Criterion for model selection among singular exponential models to take into account for the singularities in this family of models. Watanabe's book gives the general setup to address this problem but it requires some deep algebraic steps, namely resolution of singularities, that at the moment are not entirely feasible to perform in a computer. My main interest is to find ways of using the structure of these models to ease the resolution of singularities. It would also be interesting to start collaborations with experimental statisticians to study the power and relevance of these methods in real data sets.

A.5 Karwa, Vishesh

1) Learning causal Bayesian Networks:

One of the central issues in causal inference framework is learning a causal model from observational data. Causal models are generally represented in the form of a family of Bayesian Networks called Partial ancestral graphs (PAGs). PAGs are equivalence classes of Bayesian Networks with hidden variables. A very specific question is to see how can we perform learning and model selection of causal Bayesian Networks using Singular learning theory.

2) Ecological Inference

Ecological inference refers to performing statistical inference about individuals in presence of aggregated data. By nature, ecological models are non-identifiable and contain missing data. There is no well accepted solution to this problem, and most of the current solutions are applicable only to contingency tables with smaller dimension. Again, model selection is a central issue in ecological inference where tools from Singular Learning theory may help.

A.6 Kiraly, Franz

My current work as a revolves around the application of algebraic methods in Machine Learning and Data Analysis, where model selection is central problem. Thus my main questions are related to algorithmical and theoretical application of Singular Learning Theory to these fields.

Regularization. Watanabe's standard form of the log likelihood ratio is a powerful tool for fitting a parametric model to data. Intrinsic knowledge of the model space's structure and its embedded singularities is used in order to obtain this standard form. Also, the Watanabe Information Criterion (WAIC) is the canonical tool for model selection in a singular setting, which similarly applies knowledge on the model classes' structure.

A large amount of classical learning machines resp. Machine Learning algorithms uses regularization successfully in order to work around the non-convergence of the maximum likelihood estimate. It would be an interesting question to pursue whether and how the classical regularization methods relate to the methods in Singular Learning Theory, and whether new regularization strategies for loss functions and for model selection in a class of models can be derived, e.g. for hierarchical resp. structural models.

Approximation. In its heart, Singular Learning Theory relies on the assumption that the true distribution is contained in a parametric family of distributions. However, in practice, the best one can hope for is often only that the true distribution can be approximated arbitrarily closely with growing model complexity - this has been noted already on several occasions in the classical Singular Learning Theory papers. However, a more detailed analysis of the convergence behaviors of parametric model classes in terms of model complexity would probably be of interest, e.g. as it is classically done for Neural Networks in Machine Learning.

Simplicity. One of the central questions in model selection is how to find the "simplest" model. If the families of models are fixed, this is canonically done by WAIC (e.g. selecting the dimension parameter in reduced rank regression). However, the parametric families often have a nested and possibly combinatorially complicated structure (e.g. Bayesian Networks, sparsity constraints etc); also, the choice of which families of parametric models to include

bears the danger of overfitting in the model domain. While not a concrete question, this dilemma is probably interesting from a conceptual point of view.

Automatization. While Singular Learning Theory is capable of providing standard coordinates, learning rates, and model selection criteria in a canonical way, the calculations involving desingularization leading there are already very complicated for rather elementary model classes, see e.g. the series of papers of Aoyagi and Watanabe. For anyone wanting to apply those methods to real data, and subsequently for a broader Machine Learning audience, it would be very interesting to have a statistical toolbox or software library which integrates model specification, resolution of singularities, model fitting and model selection. As Algebraists, Statisticians and Machine Learners will attend the conference, this could be a perfect opportunity to discuss the limitations and chances of such a project.

A.7 Letac, Gerard

(1) Let $I_1, \dots, I_n$ be finite sets and let $\mathcal{D}$ be a family of non empty subsets of $V = \{1, \dots, n\}$ such that $D \subset D'$ and $D' \in \mathcal{D}$ implies $D \in \mathcal{D}$. The hierarchical model governed by $\mathcal{D}$ is the set of probabilities $(p(i))$ on $I = I_1 \times \dots \times I_n$ such that $p(i) > 0$ for all $i \in I$ and such that $\log p(i) = \sum_{D \in \mathcal{D}} \lambda_D(i)$ where $i \mapsto \lambda_D(i)$ does not depend on i_v when $v \notin D$. If $\mathcal{D}$ is the family of the complete subsets of a graph, the hierarchical model is said to be graphical. For instance if $n = 3$ and $\mathcal{D}_0 = \{1, 2, 3, 12, 23, 13\}$ is not graphical. If $I_v = \{0, \dots, c_v - 1\}$ and if $S(i) = \{v; i_v \neq 0\}$ introduce $J = \{i; S(i) \in \mathcal{D}\}$ and the symbol $j \lhd i$ for saying that $S(j) \subset S(i)$ and $i_{S(j)} = j_{S(j)}$. Denote by $(e_j)_{j \in J}$ the canonical basis of $\mathbb{R}^J$. The hierarchical model is actually a exponential family parameterized by $\mathbb{R}^J$ and concentrated on the polytope with extreme points $f_i = \sum_{j \lhd i} e_j$ where $i \in I$. It is generated by the uniform measure on the set of the $c_1 \dots c_n$ vectors f_i. For instance if $n = 3$ $c_1 = c_2 = 3$ and $c_3 = 2$ and $\mathcal{D} = \{1, 2, 3, 12, 23\}$ then the set J has 11 elements and the polytope in $\mathbb{R}^{11}$ has $3 \times 3 \times 2 = 18$ vertices.

(2) In general, given an open convex set C in $\mathcal{R}^d$ containing no line, its characteristic function $\mathbb{J}_C$ is the real function on C defined by

$$\mathbb{J}_C(m) = d! \mathrm{Vol}(C - m)^o = d! \int_{C^o} (1 - \langle \theta, m \rangle)^{-d-1} d\theta = \int_{\mathbb{R}^d} e^{\langle \theta, m \rangle - h_C(\theta)} d\theta$$

where $C^o = \{\theta \in \mathcal{R}^d \ ; \ \langle \theta, x \rangle \leq 1 \forall x \in C\}$ is the polar set of C and where h_C is the support function defined by $h_C(\theta) = \sup\{\langle \theta, x \rangle \ ; \ x \in C\}$

(3) A feature of the Diaconis Ylvisaker conjugate family of a natural exponential family F generated by some positive measure μ on $\mathbb{R}^d$ with Laplace transform $e^{k_\mu(\theta)}$ is this: assume here that μ has a compact convex support with a non empty interior C. If $\alpha > 0$ and if m is in C of F then $I_(m, \alpha) = \int_{\mathbb{R}^d} e^{\alpha \langle \theta, m \rangle - k_\mu(\theta))}$ is finite. Under these circumstances this is an easy exercise of calculus to prove that $\lim_{\alpha \to 0} \alpha^d I(m, \alpha) = \mathbb{J}_C(m)$. Note that this limit depends on μ only through its convex support.

(4) If we want to use the Bayesian factor method for choosing between two hierarchical models governed by $\mathcal{D}_1$ and $\mathcal{D}_2$ this Bayesian factor is expressed as

$$B(\alpha) = \frac{I_1(\frac{\alpha m_1 + t_1}{\alpha + N}, \alpha + N) I_2(m_2, \alpha)}{I_2(\frac{\alpha m_2 + t_2}{\alpha + N}, \alpha + N) I_1(m_1, \alpha)}$$

where t_1 and t_2 are in the polytopes $\overline{C_1}$ and $\overline{C_2}$ attached to the two hierarchical models and t_1 and t_2 depend on the N observations which have been done, and m_1 and m_2 are

parameters in the open convex polytopes C_1 and C_2 which have been fixed for initialization of the Bayesian procedure. The basic idea of our method is to realize that the value of α should not be important and that we can make $\alpha \to 0$ in $B(\alpha)$. If t_1 and t_2 are in C_1 and C_2, what is described above gives easily an asymptotic expression of $B(\alpha)$. If either t_1 or t_2 are on the boundary of the polytopes $\overline{C_1}$ and $\overline{C_2}$ a delicate study of the behavior of the function $\mathbb{J}_C(m)$ at the boundary of C becomes necessary.

(5) According to the interests of participants, I will give details about this asymptotic behavior of $\mathbb{J}_C(m)$ at the boundary, I will describe the properties of $\mathbb{J}_C(m)$ as a rational function when C is a general polytope, or I will make explicit calculations of $\mathbb{J}_C$ for some particularly interesting polytopes C like quadrangles in the plane, octahedron, polytopes associated to a decomposable graphical model or to a cyclic graphical model and the hierarchical model $\mathcal{D}_0$ mentioned above.

A.8 Leykin, Anton

I am interested in computing the so-called jumping numbers (in particular, the log canonical threshold) for an algebraic variety, which has applications in statistics.

A.9 Lin, Shaowei

My research interests are in algebraic statistics and computational algebraic geometry. To me, the two greatest needs in singular learning theory are as follows.

Firstly, statisticians need an intuitive geometric understanding of the complexity of a model and an example of how such an understanding can help resolve dilemmas in model selection. One way of measuring the complexity of a model is through its learning coefficient. This coefficient is in general difficult to compute. Therefore, to motivate research effort in calculating this number, we need to explore more reasons for the importance of the learning coefficient in many questions about statistical modelling and machine learning.

1a. Can we design information criteria or learning algorithms which have provably better performance and use the learning coefficient in a critical way?

1b. Can we design MCMC methods which give better approximations of the likelihood integral by using a desingularization map for the model.

1c. How do we compute the constant term in Watanabe's asymptotic expansion of the expected value of the stochastic complexity? Can we employ this constant in some information criterion?

Secondly, computational mathematicians need to come up with better methods for calculating the learning coefficient, methods which statisticians can understand and use. Currently, bounds for the learning coefficient are mainly computed using clever manipulations of the Kullback-Leibler function, or of the fiber ideal (see my PhD dissertation). There are general algorithms for finding a resolution of singularities for a function or ideal, but these algorithms are very slow on statistical problems which have high dimensional parameter spaces. Meanwhile, there are effective tools which work in nondegenerate cases, such as Varchenko's Newton polyhedra method. It would be useful to identify statistical models where this polyhedral method gives meaningful results.

2a. For directed graphical models, are most singularities simple normal crossings (SNCs)? What about tree models (see Piotr Zwiernik's work)? What about undirected models? If the singularities are SNCs, how do we compute the learning coefficient at these

points? Is the fiber ideal sos-nondegenerate? If it is, we can use the Newton polyhedra method. If not, what is a counter-example?

2b. The Newton polyhedra method works for points on the interior of the parameter space. For points on the boundary of the parameter space, how do we compute the learning coefficient? What if the boundary is a simple normal crossing? What if it is just a normal crossing? Can we extend the Newton polyhedra method to work in these cases?

We know that the learning coefficient of a model with respect to a true distribution is the minimum of learning coefficients at points in the parameter space which map to the true distribution. This raises the following question:

2c. Can we compute where this mimimum occurs on the fiber over the true distribution? Experimentally, we find that this minimum occurs at points where the local degree of the KL function (more generally, the local multiplicity of the fiber ideal) is maximized. Can we prove or disprove this conjecture?

A.10 Montufar Cuartas, Guido

Understand the geometry of Restricted Boltzmann Machines and of Deep Belief Networks.

A.11 Pericchi, Luis

I have been working in Hypothesis Testing and Model Selection, mainly under a Bayesian approach.

Some of my interest are:

1) Generalization of Objective Priors for Model Selection and hypothesis testing, like Intrinsic Priors.

2) Generalization of BIC

3) Optimal Choice of Training Samples

4) Approximations to the computation of Evidences (marginal likelihoods) and Bayes Factors

5) Bridges between Bayesian and Significance Testing Approaches. How to reconcile the disagreement?

References:

-Berger J.O. and Pericchi L.R. (1996) The Intrinsic Bayes Factor for Model Selection and Prediction. Journal of the American Statistical Association, 91, 433, p. 109-122.

-Berger J.O. and Pericchi L.R. (1996) The Intrinsic Bayes Factor for Linear Models (invited discussant: Prof. D. Dey, University of Connecticut, USA). Bayesian Statistics 5, Bernardo J.M. et al editors. Oxford University Press. Invited conference. p. 25-44

-Berger J.O. and Pericchi L.R. (2001) Objective Bayesian Model Selection. Introduction and Comparisons, in Lectures Notes of the Institute of Mathematical Statistics. ?Model Selection?, editor: P. Lahiri, pp. 135-207.

-Berger J.O. and Pericchi LR. (2004) Training samples in objective Bayesian model selection. Annals of Statistics, 32, 3, p. 841-869.

- Pericchi L.R. (2005) Model Selection and Hypothesis Testing based on Objective Probabilities and Bayes Factors. Elsevier B.V. Handbook of Statistics, vol. 25. p. 115-149.

-Pericchi, L.R. (2010) How large should be the training sample? In the book: ?Frontiers of Decision Making and Bayesian Analysis. In Honor of James O. Berger, Chen MH et al editors. Springer.P. 130-142

A.12 Petrovic, Sonja

I am attending this workshop to learn about the model selection problem at singular points. I have recently started reading Watanabe's book on the topic, but I am still not clear how a statistician might use the learning coefficients effectively in practice. I would like to see this done on an example, perhaps of the sort Mathias Drton is suggesting.

Also, I would like to see a set of examples where one can compute the learning coefficients, using Anton Leykin's code for example, for a family of Gaussian models.

Finally, I would like to learn how I can "understand" the singularities of discrete graphical models that are relevant for machine learning, for example the Boltzmann Machine.

A.13 Plummer, Martyn

My main interest is in statistical computing and model choice for Bayesian hierarchical models. In the last 10 years, the Deviance Information Criterion (DIC) has become a popular model choice criterion, largely because it is easy to calculate using Markov Chain Monte Carlo (MCMC) methods and, in particular, because it is implemented in the popular OpenBUGS software (`www.openbugs.info`).

DIC extends the Akaike Information Criterion to Bayesian hierarchical models by replacing the number of parameters p with an estimated "effective number of parameters" p_D. DIC was introduced with only a heuristic justification by Spiegelhalter et al (2002). My own work (Plummer 2008) is an attempt to establish a rigorous foundation for DIC. This work suggests that DIC is an approximation that requires certain asymptotic conditions for its validity. In particular, a necessary (but not sufficient) condition is $p_D \ll n$ where n is the sample size. Thus it seems plausible that DIC is being mis-applied to models where the asymptotic conditions do not hold. Moreover, we currently lack an *easily computable* criterion for such models.

A second issue of interest to Bayesian statisticians is how to extend DIC to models with missing data. In the Bayesian approach, missing data and unknown parameters are treated symmetrically as unobserved random variables. However, in the model choice problem, we may wish to treat the missing data as a nuisance, and the model parameters as the "focus" of interest. Finite mixture models are an important test case for this problem. An extensive survey of possible solutions was provided by Celeux et al (2006) but the question remains unresolved. Again, the problem is to find a criterion that is both theoretically sound and computationally feasible.

I hope to gain some insight into these issues from the workshop by discussing parallel developments in machine learning with other participants.

- Celeux, G., Forbes, F., Robert, C. and Titterington, M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis* **1**: 701-706.
- Plummer M. (2008) Penalized loss functions for Bayesian model comparison. *Biostatistics* **9**:523-539.
- Spiegelhalter D.J., Best, N., Carlin, B. Van der Linde A. (2002) Bayesian measures of of model complexity and fit (with discussion) *Journal of the Royal Statistical Society, Series B*, **64**: 583-639.

A.14 Ray, Surajit

My research interests are in the area of model selection, the theory and geometry of mixture models and functional data analysis. I am especially interested in challenges presented by "large magnitude", both in the dimension of data vectors and in the number of vector. Core areas of methodological research include multivariate mixtures, structural equations models, high-dimensional clustering and functional clustering. In the context of this workshop I am specifically interested in exploring two aspects of model selection in the broad area of mixture models and clustering.

The first objective is to explore the upperbound on the number of modes a two component mixture of elliptical distribution. A recent paper by Ray and Ren (2011) [To appear in Journal of Multivariate Analysis] shows that one can get as many as $D+1$ modes from a two component normal mixture in D dimensions. Is the same true for any elliptical distribution. Further I want to explore if one can obtain an upperbound on the number of modes of a mixture of k components. The conjecture is that the answer to the above questions lie in the analysis of their respective ridgeline manifold described in Ray and Lindsay (2005).

The second problem relates to the providing an inferential framework to determine the statistical significance of modal clusters (Li et. al., 2008). The test statistics based on height distributions have been recently been used for inference on modes by Comaniciu et al. (2002) and (Burman and Polonik, 2009). The steps for developing the inferential framework might involve the following steps

1. Design a test statistic for two specified clusters, based on the ratio of the heights between the saddle point and the mode with minimum height, that is $RH = \frac{f_s(x)}{f_m(x)}$.
2. For a pre specified smoothing parameter, using kernel density estimator, we can estimate this ratio by $\widehat{RH} = \frac{\hat{f}^h_s(x)}{\hat{f}^h_m(x)}$.
3. Propose the hypothesis $H_0 : RH = 1$ vs $H_A : RH < 1$ to test the significance of these two clusters. Rejection of the test will imply the existence of separate clusters.
4. Explore the sampling distribution of $\widehat{RH}$ and finally compute the critical value the test statistic to accept or reject $H_0 : RH = 1$ vs $H_A : RH < 1$
5. Use Roy's union intersection principle to decompose the complex hypothesis of "how many clusters" at a specified level and build the appropriate test statistic and the critical region to provide an answer.

But how do we arrive at the distribution of RH?

A.15 Slavkovic, Aleksandra

1. Ecological Inference

Ecological inference refers to performing statistical inference about individuals in presence of aggregated data (i.e., only partial information is available). By nature, ecological models are non-identifiable and contain missing data. There is no well accepted solution to this problem, and most of the current solutions are applicable only to contingency tables with smaller dimension. There are links to latent class models, but this link has not been explored and the model selection is problematic. These models are also hierarchical in nature relying on Bayesian inference, again facing issue with model selection. We have begun some exploratory analysis of understanding the geometry of these models.

2. Causal inference with observational data

Linked to the above ecological inference problems, is causal inference with observational data. One of the central issues in causal inference framework is learning a causal model from observational data. Different frameworks have been proposed, namely Potential Outcome and Causal Diagrams. The latter are typically tied to Causal models that are generally represented in the form of a family of Bayesian Networks called Partial ancestral graphs (PAGs). PAGs are equivalence classes of Bayesian Networks with hidden variables. A very specific question is to see how can we perform learning and model selection of causal Bayesian Networks. But more generally, can we utilize tools from algebraic geometry and singular learning theory to establish equivalences between assumptions of these two frameworks. In a paper by Karwa, Slavkovic and Donnell (2011), we discuss some of these issues but from applied perspective (arxiv.org/pdf/1107.4855).

Generalized Estimating Equations (GEEs) have not been considered in either of these two settings, and their semi-parametric nature may help with model fit and selection. Furthermore, I am interested in possible developments of model selection criteria that take into consideration finite sample properties.

A.16 Sturmfels, Bernd

I am interested in algorithms in algebraic geometry and their applications in wide range of contexts, including those in statistics. The analysis of singularities for marginal likelihood integrals and the resulting refined information criteria tie in very nicely with topics of considerable interest to algebraic geometers, such as the log-canonical threshold. I am optimistic that, during the workshop week, we can achieve much progress on the computation of such quantities for relevant models. One such interesting model is the restricted Boltzmann machine.

A.17 Sullivant, Seth

I am attending this workshop to learn more about this topic. My impression is that most of the work in this area has been focused on the case of models for discrete random variables (e.g. latent variable discrete graphical models). I am curious as to what can be done for gaussian graphical models. There are many families of latent variable gaussian graphical models where explicit defining equations are known, the equations are determinantal constraints on a symmetric covariance matrix, and it should be possible to analyze the singularities in some cases.

A.18 Watanabe, Sumio

Introduction

This short article introduces research themes which are discussed in the workshop, "Singular learning theory: algebraic geometry and model selection", held at American Institute of Mathematics, December 12-16, 2011.

From Statistics to Algebraic Geometry

Many statistical models which have hidden variables, hierarchical structures, or submodules are not regular statistical models, because neither the map from the parameter to the probability distribution is one-to-one nor Fisher information matrix is positive definite. In model selection or statistical hypothesis test for such statistical models, the likelihood

function can not be approximated by any quadratic form. Therefore, neither the log likelihood ratio is subject to the χ^2 distribution nor AIC, BIC, MDL, or DIC has theoretical foundation. Conventional statistical asymptotic theories do not hold.

In statistical model evaluation process, we have three important random variables. Let $X_1, X_2, ..., X_n$ be random variables which are independently subject to the probability density function $q(x)$. For a statistical model $p(x|w)$ and a prior $\varphi(w)$, the free energy or the minus Bayes log marginal is defined by

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

The generalization error G and the training error T are also defined by

$$\begin{aligned} G &= -\int q(x) \log \mathbb{E}_w[p(x|w)]dx, \\ T &= -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{E}_w[p(X_i|w)], \end{aligned}$$

where $\mathbb{E}_w[\ ]$ denotes the expectation value over the posterior distribution with the inverse temperature β. These three variables are invariant under an analytic transform $w = g(u)$ by

$$\begin{aligned} w &= g(u) \\ p(x|w)dx &= p(x|g(u))dx \\ \varphi(w)dw &= \varphi(g(u))|g'(u)|du \end{aligned}$$

where $|g'(u)|$ is the Jacobian determinant. Hence They are understood as the generators of the birational invariants.

In fact, the first nontrivial term in the asymptotic expansion of F is equal to $\lambda \log n$ where λ is the well-known birational invariant, *the real log canonical threshold*. Moreover, the first nontrivial term of $\mathbb{E}[G]$ is equal to $\{(\lambda - \nu)/\beta + \nu\}/n$ where $\nu = \nu(\beta)$ is the new birational invariants, *the singular fluctuation* that satisfies

$$\nu(\beta) = \lim_{n\to\infty} \frac{\beta}{2}\mathbb{E}[V],$$

where

$$V = \sum_{i=1}^{n} \{\mathbb{E}_w[(\log p(X_i|w))^2] - \mathbb{E}_w[\log p(X_i|w)]^2\}.$$

Therefore the singular fluctuation shows a kind of variance of the log likelihood function near singularities.

In regular statistical models, $\lambda = \nu = d/2$, where d is the dimension of the parameter, which is the mathematical foundation of BIC and AIC. However, in general, they are different from $d/2$. We would like to ask what mathematical properties such birational invariants represent. Also we want to know the mathematical foundation on which such problem will be resolved.

This issue has connections to large fields in mathematics, algebraic geometry, algebraic analysis, and commutative ring theory. In fact, resolution of singularities, Bernstein-Sato polynomial, toric modification using Newton diagram, Jet-scheme analysis and so on. We

expect that mathematical questions in singular statistics make us open the new mathematical field where algebraic variety-valued random variables are studied.

From Algebraic Geometry to Statistics

On the other hand, there are at least three different applications of algebraic geometry to statistics.

The first application of algebraic geometry is the direct evaluation of several random variables. For example, if we know the real log canonical threshold and singular fluctuation, then we can evaluate how appropriate the statistical model and the prior compared to the numerical values of the free energy or the generalization error. These are the generalizations of BIC and AIC in regular models to general models.

The second application is the evaluation of the Markov chain Monte Calro (MCMC) methods. In order to approximate the posterior distribution, MCMC method is often necessary. However, it is still difficult to evaluate the MCMC simulations. For a given set of a true distribution, a statistical model, and a prior, we can evaluate the accuracy of the MCMC simulation if we know the real log canonical threshold and the singular fluctuation.

The last application is to find the mathematical law in statistics based on algebraic geometry. For example, based on algebraic geometrical method, we can prove that there are universal relation among G, T, and V

$$\mathbb{E}[G] = \mathbb{E}[T] + \frac{\beta}{n}\mathbb{E}[V] + o(\frac{1}{n}).$$

This relation holds for both regular and singular cases, hence is very useful to estimate the generalization error from the training error. Recently, we found that this equation is asymptotically equivalent to the cross-validation.

Algebraic Geometry and Statistics

To study an algebraic variety V, we need the ideal $\mathbb{I}(V)$. If V is the set of true parameters in statistics, then the log density ratio function satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i|w)} \in \mathbb{I}(V).$$

Also its expectation satisfies

$$\int q(x) \log \frac{q(x)}{p(x|w)} dx \in \mathbb{I}(V).$$

In regular statistical theory, only maximal ideal

$$\mathbb{I}(V) = \langle w_1, w_2, ..., w_d \rangle$$

was studied because the log likelihood function can be approximated by a quadratic form. However, in general statistical theory, general ideals are necessary. For example, in a regression model,

$$Y = a\tanh(bx) + c\tanh(dx) + \text{noise},$$

for the case that the true distribution is $Y = 0 + \text{noise}$, the ideal is

$$\mathbb{I}(V) = \langle ab + cd, ab^3 + cd^3 \rangle.$$

This fact shows that

"To estimate the structure of the true distribution, algebraic geometry is necessary."

We expect that a lot of algebraic geometrical studies are essential to future statistics and that new birational invariants are found in statistics.

A.19 Xi, Jing

I am interested in studying contingency tables whose cells are distributed according to a hierarchical loglinear model. I have worked on estimating the number of multi-way contingency tables as well as multi-way zero-one contingency tables via sequential importance sampling procedures (see more details at xia,xib).

In the workshop I am interested in computational experiment on the Diaconis-Ylvisaker conjugate prior D-Y with the software `LattE Integrale`. For simulation study in my previous work, I have used the older version of `LattE` lattesoft. Thus it is familiar for me to use this software. The Diaconis-Ylvisaker conjugate prior is the form of

$$I(m,\alpha)^{-1}L(\theta)^{-\alpha}\exp(\alpha<\theta,m>)d\theta$$

where L is its Laplace transform, m and α are hyperparameters and $I(m,\alpha)$ is the normalization constant. [Massam] called this prior as the generalized hyper Dirichlet. In this workshop I will focus on the computation of $I(m,\alpha)$. This quantity is needed to be computed when doing a Bayesian model search. $I(m,\alpha)$ is the form of

$$I(m,\alpha)=\int_{\mathbf{R}^{|J|}}L(\theta)^{-\alpha}\exp(\alpha<\theta,m>)d\theta,$$

where J is a subset of the index set of cells I for contingency tables defined by a certain model. This is the integration over the $|J|$-dimensional unit cube, of the power $-\alpha$ of a polynomial of degree at most $|J|$, multiplied by a product of $|J|$ independent beta densities. [Massam] showed this integration as the form of rational function (Equation (15)). Therefore it would be interesting to compute these rational functions with the software `LattE Integrale` latte1.

Bibliography

[Massam] G. Letac and H. Massam, *Bayes factors and the geometry of discrete hierarchical loglinear models*, 2011, notes = Available at http://arxiv.org/abs/1103.5381 [D-Y] Diaconis, P. and Ylvisaker, D., 1979, *Conjugate priors for exponential families*, Ann. Statist., 7, 269–281 [xia] Author = J. Xi and R. Yoshida, Notes = Preprint. Available at http://arxiv.org/abs/1108.5939, Estimating the number of 0-1 multi-way tables via sequential importance sampling, 2011

[xib] Author = J. Xi and S. Wei and F. Zhou and D. Haws and R. Yoshida, Notes = Preprint. Available at http://arxiv.org/abs/1108.2311, Semigroups and sequential importance sampling for multiway tables and beyond, 2011

[latte1] notes= Available at http://front.math.ucdavis.edu/1108.0117, J. De Loera and B. Dutra and M. Koeppe and S. Moreinis and G. Pinto and J. Wu, 2011, *Software for Exact Integration of Polynomials over Polyhedra*

[lattesoft] Author = De Loera, J. A. and Haws, D. and Hemmecke, R. and Huggins, P. and Tauzer, J. and Yoshida, R., Date-Added = 2011-07-29 15:28:41 -0400, Date-Modified = 2011-10-24 18:59:42 -0400, Software and user's guide for LattE v.1.1, Url = www.math.ucdavis.edu/∼latte, 2005, Bdsk-Url-1 = www.math.ucdavis.edu/∼latte

A.20 Zhang, Yongli

In the area of model selection I am particularly interested in the following questions:

- How to identify the true model as the number of variables p is much greater than the sample size n? Especially, the n-related criterion like BIC is inadequate in high dimensional data;
- The estimation of risk (MSE) as the model selection uncertainty is taken into account.

Some progress has been made in the above two areas and the results are presented in Zhang and Shen (2010a) and Zhang and Shen (2010b). Concerning singular learning theory I want to touch on the following two questions:

- The Hidden Markov Model is widely used in econometric time series modeling, but the number of states is often given a priori without any empirical evidence. How to select the number of states is worth future input.
- As is well known LASSO does not work very well as there exists high correlation between predictors in the true model. Could WAIC play a role in high dimensional data learning?

Bibliography

[zhang] Zhang, Y. and Shen, X. (2010a). Model selection procedure for high-dimensional data. *Statistical Analysis and Data Mining*, **3**, 350-358

[zhang] Zhang, Y. and Shen, X. (2010b). Optimal data perturbation size in modeling procedure risk estimation. *Manuscript*

A.21 Zwiernik, Piotr

My research focuses on graphical models with hidden variables. In particular I am interested in models induced by directed trees such that all the inner nodes represent hidden variables. Some familiar examples are hidden Markov models and general Markov models used in phylogenetics. In the context of singular learning theory these models are particularly interesting and have many links to algebraic geometry and combinatorics. The singularities which arise are simple normal crossings with nice poset structure. My recent work on this topic will appear in Journal of Machine Learning Research. This paper was an extension of a previous result of Rusakov and Geiger (2005). I obtained the generalization of the BIC formula in the case when the limit of the likelihood function for large sample sizes is maximized over a singular subset of the parameter space. The main idea was to use the geometric understanding of possible fibers of the parametrization.

I would be interested in studying similar problems for related classes of models. However, in addition, I would like to understand if there is any efficient way of dealing with points on the boundary of the parameter space at least in some simple cases.