Generalized BIC for Singular Models Factoring through Regular Models

Shaowei Lin

13 Dec 2010

Abstract

The Bayesian Information Criterion (BIC) is an important tool for model selection, but its use is limited only to regular models. Recently, Watanabe [5] generalized the BIC to singular models using ideas from algebraic geometry. In this paper, we compute this generalized BIC for singular models factoring through regular models, relating it to the real log canonical threshold of polynomial fiber ideals.

1 Introduction

In statistical learning theory, the Bayesian Information Criterion (BIC) is an important tool for model selection. It is defined by

$$-\log L_0 + \frac{d}{2} \log N$$

where L_0 is the maximum likelihood, d the dimension of the model, and N the sample size. The BIC is an asymptotic result derived for identifiable models whose Fisher information matrix is positive definite [6, 5]. Such models are called *regular*; otherwise, they are *singular*. In 2001, Watanabe showed that there are singular models for which the BIC is not asymptotically equal to the Bayesian marginal likelihood [4]. He proved that under mild conditions, the BIC can be generalized to singular models with the formula

$$-\log L_0 + \lambda \log N - (\theta - 1) \log \log N,$$

where λ is the smallest pole of the zeta function

$$\zeta(z) = \int_{\Omega} K(\omega)^{-z} \varphi(\omega) d\omega, \quad z \in \mathbb{C}$$

and θ the multiplicity of this pole. Here, Ω is the model parameter space, $\varphi(\omega)$ is a prior belief on Ω, and $K(\omega)$ the Kullback-Leibler distance to the maximum likelihood distribution. In algebraic geometry, the pair (λ, θ) is called the *real log canonical threshold* of $K(\omega)$. To compute (λ, θ), we need to find a resolution of singularities for the analytic function $K(\omega)$, that is, a change of variables $\omega(w)$ such that $K(\omega(w))$ is a monomial function of w. By a famous theorem of Hironaka [2], such resolutions always exist, but are in practice difficult to find. Nonetheless, Watanabe's work uncovered interesting connections between statistical learning theory and algebraic geometry.

In a previous paper [3], we studied how resolutions of $K(\omega)$ may be found for models on a finite discrete space $\{1, 2, \ldots, k\}$. If the model is parametrized by polynomials $p_1(\omega), \ldots, p_k(\omega)$, we showed that the pair $(2\lambda, \theta)$ is the real log canonical threshold (RLCT) of the polynomial *fiber ideal*

$$\langle p(\omega) - q \rangle := \langle p_1(\omega) - q_1, \ldots, p_k(\omega) - q_k \rangle$$

where $q_1, \ldots, q_k$ are relative frequencies coming from the data. To derive the RLCT, we monomialize the fiber ideal. Computationally, monomializing a polynomial ideal is easier than monomializing an analytic function. Moreover, in nondegenerate cases, we can compute the RLCT of an ideal using a geometric-combinatorial method involving Newton polyhedra.

In this paper, we continue this vein of study and describe fiber ideals for non-discrete models. In particular, we explore models which factor through regular models. Consider an regular model $\mathcal{M}_1$ over a parameter space U with probability density function $p(x|u, \mathcal{M}_1), u \in U$. We say that a model $\mathcal{M}_2$ over some space Ω factors through $\mathcal{M}_1$ if $p(x|\omega, \mathcal{M}_2) = p(x|u(\omega), \mathcal{M}_1)$ for some function $u(\omega)$. Such models occur frequently in statistical learning, for instance, multivariate Gaussian mixed-graph models $\mathcal{N}(0, \Sigma)$ where the covariance matrix Σ is parametrized by

$$\Sigma = (I - \Lambda)^{-T} \begin{pmatrix} K^{-1} & 0 \\ 0 & \Phi \end{pmatrix} (I - \Lambda)^{-1}.$$

for some parameter matrices Λ, K and Φ. We will show that if the maximum likelihood estimate $\hat{u}$ in $\mathcal{M}_1$ of the data also lies in $\mathcal{M}_2$, then the Bayesian marginal likelihood can be approximated using the RLCT of the fiber ideal

$$\langle u(\omega) - \hat{u} \rangle.$$

We end by studying this result for exponential families and, specifically, for multivariate Gaussian models.

2 Asymptotics of Laplace Integrals

In this section, we give an overview of the asymptotics of Laplace integrals

$$Z(N) = \int_\Omega e^{-NK(\omega)} \varphi(\omega) d\omega$$

where $K(\omega)$ and $\varphi(\omega)$ are analytic functions over some space

$$\Omega = \{\omega \in \mathbb{R}^d \mid g_1(\omega) \geq 0, \ldots, g_l(\omega) \geq 0\}$$

defined by analytic inequalities. Let $K_0 = \inf_{\omega \in \Omega} K(\omega)$. First, observe that

$$Z(N) = e^{-NK_0} \int_\Omega e^{-N(K(\omega)-K_0)} \varphi(\omega) d\omega$$

so we will now assume that $\inf K(\omega) = 0$. Then, such integrals have asymptotic expansions [1, 3, 5]

$$Z(N) \;\approx\; \sum_\alpha \sum_{i=1}^d c_{\alpha,i} N^{-\alpha} (\log N)^{i-1}, \quad N \to \infty$$

where the sum is over positive rationals α and reals $c_{\alpha,i}$. The leading term

$$Z(N) \;\propto\; N^{-\lambda} (\log N)^{\theta-1}, \quad N \to \infty$$

is given by the smallest pole λ and multiplicity θ of the zeta function

$$\zeta(z) = \int_\Omega K(\omega)^{-z} \varphi(\omega) d\omega, \quad z \in \mathbb{C}.$$

We denote the pair (λ, θ) by $\mathrm{RLCT}_\Omega(K; \varphi)$. We order such pairs by $(\lambda_1, \theta_1) > (\lambda_2, \theta_2)$ if $\lambda_1 > \lambda_2$ or if $\lambda_1 = \lambda_2$ and $\theta_1 < \theta_2$.

Following [3], we define RLCTs for ideals. Suppose $I = \langle f_1(\omega), \ldots, f_r(\omega)\rangle$ is a finitely-generated ideal in the ring of functions real-analytic over Ω. Let $\mathrm{RLCT}_\Omega(I;\varphi)$ denote the smallest pole and multiplicity of the zeta function

$$\zeta(z) = \int_\Omega (f_1(\omega)^2 + \cdots + f_r(\omega)^2)^{-z/2}\varphi(\omega)d\omega, \quad z \in \mathbb{C}.$$

One can show that this definition is independent of the choice of generators $f_1, \ldots, f_r$ for I. Furthermore, this RLCT is the minimum

$$\mathrm{RLCT}_\Omega(I;\varphi) = \min_{\omega \in \mathcal{V}_\Omega(I)} \mathrm{RLCT}_{\Omega_\omega}(I;\varphi)$$

of local RLCTs where Ω_ω is a sufficiently small neighborhood of ω in Ω and $\mathcal{V}_\Omega(I)$ is the variety $\{\omega \in \Omega \mid f(\omega) = 0 \ \forall f \in I\}$.

We now describe a geometric-combinatorial tool for computing the local RLCT of the ideal I at an interior point $\omega_0 \in \Omega$. After a translation, we may assume that ω_0 is the origin. Recall that I is an ideal of functions analytic at the origin, and $0 \in \mathcal{V}(I)$. Given $f(\omega) \in I$, let $f(\omega) = \sum_\alpha c_\alpha \omega^\alpha$ be its power series expansion where $\omega = (\omega_1, \ldots, \omega_d)$ and $\alpha = (\alpha_1, \ldots, \alpha_d)$. Define its *Newton polyhedron* $\mathcal{P}(I) \subset \mathbb{R}^d$ to be the convex hull

$$\mathcal{P}(I) = \mathrm{conv}\,\{\alpha + \alpha' \mid \alpha \in \mathbb{Z}^d_{\geq 0}, \alpha' \in \mathbb{R}^d_{\geq 0}, \sum_\alpha c_\alpha \omega^\alpha \in I, c_\alpha \neq 0\}.$$

Given a face γ of $\mathcal{P}(I)$ and $f(\omega) = \sum_\alpha c_\alpha \omega^\alpha$, define the *face polynomial*

$$f_\gamma = \sum_{\alpha \in \gamma} c_\alpha \omega^\alpha$$

and the *face ideal* $I_\gamma = \langle f_{1\gamma}, \ldots, f_{r\gamma}\rangle$ where $f_1, \ldots, f_r$ generate I. This ideal is independent of the choice of generators. We say that the ideal I is *sum-of-squares(sos)-nondegenerate* if and only if for all compact faces γ of $\mathcal{P}(I)$, the variety $\mathcal{V}(I_\gamma)$ over the torus $(\mathbb{R}^*)^d$ is empty; otherwise, I is *sos-degenerate.*

Theorem 2.1 ([3, Thm 5.7])**.** *Let I be a finitely generated ideal of functions analytic at the origin and let $\varphi(\omega) = \sum_\tau c_\tau \omega^\tau$. Then,*

$$\mathrm{RLCT}_0(I;\varphi) \leq \min\{(1/l_\tau, \theta_\tau) \mid \tau \in \mathbb{Z}^d_{\geq 0}, c_\tau \neq 0\}$$

where each l_τ is the τ-distance of $\mathcal{P}(I)$ and θ_τ its multiplicity. Equality occurs when I is monomial or, more generally, sos-nondegenerate.

3 Factoring through Regular Models

In this section, we prove our main theorem on approximating the Bayesian marginal likelihood for models which factor through regular models. First, given a vector or matrix M of analytic functions, let $\langle M\rangle$ denote the ideal generated by the entries of M. We begin with the following proposition.

Proposition 3.1. *Let $u : \Omega \to U$ be analytic at $\omega_0 \in \Omega$ and $K : U \to \mathbb{R}$ be analytic at $u_0 = u(\omega_0) \in U$. Suppose $K(u_0) = \nabla K(u_0) = 0$ and the Hessian of $K(u)$ is full rank at u_0. Then, for all $\varphi(\omega)$ analytic at ω_0,*

$$\mathrm{RLCT}_{\omega_0}(K(u(\omega)); \varphi) = (\lambda, \theta)$$

where $(2\lambda, \theta) = \mathrm{RLCT}_{\omega_0}(\langle u(\omega) - u_0\rangle; \varphi)$.

Proof. Without loss of generality, let us assume that u_0 is the origin. Let d be the dimension of U. Then, there exists a choice of local coordinates $u_1, \ldots, u_d$ such that the power series expansion of $K(u)$ is $u_1^2+\cdots+u_d^2+O(u^3)$. Moreover, there is a sufficiently small neighborhood $\tilde{U} \subset U$ of the origin such that

$$c(u_1^2 + \cdots + u_d^2) \leq K(u) \leq C(u_1^2 + \cdots + u_d^2), \quad \forall\, u \in \tilde{U}$$

for some positive constants c and C. Now, since $u : \Omega \to U$ is continuous at ω_0, there exists some neighborhood $\tilde{\Omega} \subset \Omega$ of ω_0 such that $u(\tilde{\Omega}) \subset \tilde{U}$. Thus,

$$c(u_1(\omega)^2 + \cdots + u_d(\omega)^2) \leq K(u) \leq C(u_1(\omega)^2 + \cdots + u_d(\omega)^2), \quad \forall\, \omega \in \tilde{\Omega}$$

and so by [5, Remark 7.2],

$$\mathrm{RLCT}_{\omega_0}(K(u(\omega)); \varphi) = \mathrm{RLCT}_{\omega_0}(u_1(\omega)^2 + \cdots + u_d(\omega)^2; \varphi) = (\lambda, \theta).$$

Finally, by definition, $(2\lambda, \theta) = \mathrm{RLCT}_{\omega_0}(\langle u(\omega) - u_0\rangle; \varphi)$. □

Next, let us define regular models. A model $\mathcal{M}$ with parameter space U is said to be *identifiable* if

$$p(x|u, \mathcal{M}) = p(x|u', \mathcal{M})\ \forall x \quad \Rightarrow \quad u = u'.$$

Given $u_0 \in U$, we define the Kullback-Leibler distance to $p(x|u_0)$ as

$$K(u) = K(u\|u_0) = \int p(x|u_0) \log \frac{p(x|u_0)}{p(x|u)} dx.$$

The Fisher information matrix $I(u_0)$ is then the Hessian matrix of $K(u\|u_0)$ evaluated at u_0, i.e.

$$I_{jk}(u_0) = \frac{\partial^2 K}{\partial u_j \partial u_k}(u_0).$$

We say that $\mathcal{M}$ is regular if $\mathcal{M}$ is identifiable and its Fisher information matrix $I(u)$ is positive definite at all $u \in U$.

Now, given a model $\mathcal{M}$ (not necessarily regular) and independent identically-distributed samples $x_1, \ldots, x_N$, the likelihood of the data given u is

$$L(u) = \prod_{i=1}^{N} p(x_i|u)$$

while the Bayesian marginal likelihood of $\mathcal{M}$ is

$$\int_U L(u)\varphi(u)du = L_0 \int_U e^{-N\hat{K}(u)}\varphi(u)du$$

where $\varphi(u)$ is a prior on U. Here, L_0 is the maximum likelihood and

$$\hat{K}(u) = -\frac{1}{N}\left(\log L_0 + \sum_{i=1}^{N} \log p(x_i|u)\right).$$

If there is a unique distribution $q(x)$ giving the maximum likelihood, then $\hat{K}(u)$ is the empirical Kullback-Leibler distance to $q(x)$. For large sample sizes N, we approximate the Bayesian marginal likelihood with the integral

$$Z(N) = L_0 \int_U e^{-NK(u)}\varphi(u)du \tag{1}$$

where $K(u)$ is the classical Kullback-Leibler distance of $p(x|u)$ to $q(x)$.

Theorem 3.2. *Let M_1 and M_2 be models with parameter spaces U and Ω respectively such that M_1 is regular and M_2 factors though $\mathcal{M}_1$ via $u : \Omega \to U$. Let $x_1, \ldots, x_N$ be independent identically-distributed samples whose maximum likelihood estimate in $\mathcal{M}_1$ is $\hat{u} \in U$. If $u^{-1}(\hat{u}) \subset \Omega$ is nonempty, then asymptotically as $N \to \infty$,*

$$-\log Z(N) \approx -\log L_0 + \lambda \log N - (\theta - 1)\log\log N$$

where $(2\lambda, \theta) = \mathrm{RLCT}_\Omega(I; \varphi)$ and I is the fiber ideal

$$I = \langle u(\omega) - \hat{u} \rangle.$$

Proof. Direct application of Proposition 3.1. □

4 Exponential Families

Let $\mathcal{M}$ be an exponential family

$$p(x|\eta) = h(x)\exp\{\eta^T T(x) - A(\eta)\}$$

for some functions $h(x), T(x)$ and $A(\eta)$. Given independent and identically-distributed samples $x_1, \ldots, x_N$, the empirical Kullback-Leibler distance is

$$\hat{K}(\eta) = (A(\eta) - \eta^T \hat{\mu}) - (A(\hat{\eta}) - \hat{\eta}^T \hat{\mu})$$

where

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} T(x_i)$$

and $\hat{\eta}$ is the maximum likelihood estimate satisfying $\nabla A(\hat{\eta}) = \hat{\mu}$. Amazingly, $\hat{K}(\eta)$ coincides with the classical Kullback-Leibler distance $K(\eta)$ of $p(x|\eta)$ to $p(x|\hat{\eta})$, so the approximation in (1) is actually exact. Therefore, the integral $Z(N)$ in Theorem 3.2 is the Bayesian marginal likelihood. We now apply the theorem to some well-known regular exponential families.

4.1 Discrete Models

A discrete model with state space $\{1, 2, \ldots, k\}$ parametrized by state probabilities $p_1, \ldots, p_k$ is an exponential family $p(x|\eta) = \exp\{\eta^T T(x)\}$ with

$$\eta = (\log p_1, \ldots, \log p_k), \quad T(x) = (\delta(x-1), \ldots, \delta(x-k)).$$

Discrete models are regular, so Theorem 3.2 gives us the results of [3].

4.2 Multivariate Gaussian Models

The probability density function of $\mathcal{N}(\mu, \Sigma)$ is given by

$$\begin{aligned} p(x|\mu,\Sigma) &= \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}}\exp\{-\frac{1}{2}\langle (x-\mu)(x-\mu)^T, \Sigma^{-1}\rangle\} \\ &= \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}}\exp\{-\frac{1}{2}\langle xx^T, \Sigma^{-1}\rangle + \langle x, \Sigma^{-1}\mu\rangle - \frac{1}{2}\langle \mu\mu^T, \Sigma^{-1}\rangle\} \end{aligned}$$

This can be expressed as an exponential family with

$$\begin{aligned} h(x) &= \frac{1}{(2\pi)^{k/2}} \\ \eta(\Sigma, \mu) &= (\Sigma^{-1}, \Sigma^{-1}\mu) \\ T(x) &= (-\frac{1}{2}xx^T, x) \\ A(\Sigma, \mu) &= \frac{1}{2}\left(\log|\Sigma| + \langle \mu\mu^T, \Sigma^{-1}\rangle\right). \end{aligned}$$

Multivariate Gaussian models are regular, so we may apply Theorem 3.2. Let $\mathcal{N}(\mu(\omega), \Sigma(\omega))$ be a model parametrized by $\omega \in \Omega$. Given data $x_1, \ldots, x_N$, suppose the empirical mean and covariance matrix

$$\begin{aligned} \hat{\mu} &= \frac{1}{N}\sum_{i=1}^{N} x_i \\ \hat{\Sigma} &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})(x_i - \hat{\mu})^T \end{aligned}$$

lie in the model, i.e. there exists $\omega_0 \in \Omega$ such that $\mu(\omega_0) = \hat{\mu}$ and $\Sigma(\omega_0) = \hat{\Sigma}$. Then, the Bayesian marginal likelihood $Z(N)$ can be approximated by

$$-\log Z(N) \approx -\log L_0 + \lambda \log N - (\theta - 1)\log\log N$$

where $(2\lambda, \theta)$ is the RLCT of the fiber ideal

$$\langle \Sigma(\omega) - \hat{\Sigma}, \mu(\omega) - \hat{\mu}\rangle.$$

Acknowledgements I would like to thank Michael Jordan and Martin Wainwright for the opportunity to pursue this project in their statistical learning theory course. My appreciation also goes to Mathias Drton and Sumio Watanabe for many useful discussions.

References

[1] V. I. Arnol'd, S. M. Guseĭn-Zade and A. N. Varchenko: *Singularities of Differentiable Maps*, Vol. II, Birkhäuser, Boston, 1985.

[2] H. Hironaka: Resolution of singularities of an algebraic variety over a field of characteristic zero I, II, *Ann. of Math.* (2) **79** (1964) 109–203.

[3] S. Lin: Asymptotic Approximation of Marginal Likelihood Integrals, preprint `arXiv:1003.5338` (2010).

[4] S. Watanabe: Algebraic analysis for nonidentifiable learning machines, *Neural Computation* **13** (2001) 899–933.

[5] S. Watanabe: *Algebraic Geometry and Statistical Learning Theory*, Cambridge Monographs on Applied and Computational Mathematics **25**, Cambridge University Press, Cambridge, 2009.

[6] G. Schwarz: Estimating the Dimension of a Model, *Annals of Statistics* **6** (1978) 461–464.