A Tutorial on Schwarz' paper.

Dmitry Rusakov

Abstract

Schwarz [Schwarz, 1978] addresses the problem of selecting one of a number of statistical models of different dimensions. The problem is treated in Bayesian framework by evaluating the leading terms in asymptotic expansion of posterior model probability. The resulting asymptotic criterion for model selection is now termed the *Bayesian Information Criterion* (BIC) and is widely used in practice.

This paper provides a tutorial on the original Schwarz' paper [Schwarz, 1978]. It gives the introduction on the topics essential to the understanding of the asymptotic result, e.g. sufficient statistics, exponential distributions and Laplace approximation of integrals. The asymptotic expansion of posterior model probability is rigorously derived in great detail and a number of extension are introduced and discussed.

1 Introduction

Statisticians are often faced with the problem of choosing the appropriate dimensionality of a model that will fit a given set of observations. One example of such problem is the choice of structure in learning of Bayesian networks [Heckerman, 1995, Heckerman et al., 1995, Cooper and Herskovits, 1992]. In such cases the maximum likelihood principle invariably leads the model of highest possible dimension. Therefore it cannot be the right formalization of the intuitive notion of choosing the "right" model.

In Bayesian approach to model selection, the model M is chosen according to the maximum posteriori probability given the observed data D:

$$P(M|D) \propto P(M,D) = P(M)P(D|M) = P(M) \int P(D|M,\theta)P(\theta|M)d\theta \tag{1}$$

where θ denotes the model parameters.

Schwarz presented an asymptotic expansion of $P(M,D)$ using the Laplace approximation of integral in Equation 1. The result is obtained for observations that come from linear exponential family (LEF) and nowhere vanishing a priori parameter distributions $P(\theta|M)$.

Given a data D and given a model M, described by maximum likelihood function ML and dimension k, the asymptotic expansion of $\ln P(D|M)$ is given by

$$\ln P(D|M) = \ln ML(D|M) - \frac{1}{2}k \ln n + O(1). \tag{2}$$

Later this result has been extended by Haughton [Haughton, 1988] for curved exponential families (CEFs) and not fixed averaged sufficient statistics of D. She also shows the next term of asymptotic expansion and proves that BIC model selection criterion based on the presented expansion (2) gives the correct model with probabilities converging to 1 as $n \to \infty$.

This tutorial derives in great detail the results in [Schwarz, 1978] providing all the necessary mathematical and statistical background in Section 2. It is hoped that this careful derivation will enable us to extend Schwarz' results to stratified exponential models [Geiger et al., 2001]. Recently, a very encouraging results in this direction were achieved by Watanabe [Watanabe, 2000].

2 Preliminaries

This section describes some basic facts that are necessary for understanding the results presented in the following sections. These include an introduction to basic statistic concepts, such as sufficient statistics and exponential family of distributions, and a number of topics in real analysis. As a matter of convention, from now on we assume that all vectors used in derivations are column vectors and we let $a \cdot b$ denote the scalar (dot) product of vectors a and b.

2.1 Sufficient Statistics

Assume that a random variable X takes values in the sample space S (possibly vector valued) and comes from either a discrete distribution or a continuous distribution, and we shall let $f(x|\theta)$ denote the g.p.d.f.of this distribution. We use the term *generalized probability density function* (abbreviated g.p.d.f.) to denote a function that may be either a p.d.f.(probability density function) or a p.f.(probability function) ([DeGroot, 1970] page 19). We also assume that the unknown value of θ belongs to some specified parameter space Θ.

Definition. Any real-valued function $T = r(X_1, \ldots, X_n)$ of the observations in the random sample is called a *statistics* ([DeGroot, 1986] page 356).

For example, the statistics r_1 may be the average of $X_1, \ldots, X_n$, e.g. $r_1(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$, or empirical variance $r_2(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]^2$. For observations that come from normal distribution, the empirical mean (r_1) and variance (r_2) are the only statistics that are needed to estimate the distribution parameters. Such statistics are called *sufficient*.

For any prior g.p.d.f. ξ of Θ and any observed value $x \in S$, let $\xi(\cdot|x)$ denote the posterior g.p.d.f. of Θ. For simplicity, it will be assumed that for every value of $x \in S$ and every prior g.p.d.f. ξ, the posterior g.p.d.f. $\xi(\cdot|x)$ exists and is specified by Bayes theorem ([DeGroot, 1970] pages $155 - 156$).

Definition. A statistics T is a *sufficient statistics* for the family of g.p.d.f.'s $\{f(\cdot|\theta)|\theta \in \Theta\}$ if $\xi(\cdot|x_1) = \xi(\cdot|x_2)$ for any prior g.p.d.f. ξ and any $x_1, x_2 \in S$, s.t. $T(x_1) = T(x_2)$.

Loosely speaking, a sufficient statistics is a statistics s.t. for any prior distribution on Θ, its posterior distribution depends on the observed value of X only through T.

The following theorem, which is known as a *factorization criterion*, provides an easy way of recognizing sufficient statistics.

Theorem 1 (The Factorization Criterion ([DeGroot, 1970] page 156)) *A statistics T is sufficient for a family of g.p.d.f.'s $\{f(\cdot|\theta)|\theta \in \Theta\}$ if and only if $f(x|w)$ can be factored as follows for all values $x \in S$ and $\theta \in \Theta$:*

$$f(x|\theta) = u(x)v[T(x), \theta] \tag{3}$$

where u is a positive function that does not depend on θ and v is nonnegative function that depends on x only through $T(x)$.

In the normal distribution example, the empirical mean and variance statistics (r_1, r_2) can be seen to be sufficient from the above theorem, since we can factorize the $f(X_1, \ldots, X_n|\mu, \rho^2)$ as

$$\begin{aligned} f(X_1, \ldots, X_n|\mu, \sigma^2) \propto e^{-\frac{1}{2\sigma}\sum_{i=1}^{n}(X_i-\mu)^2} &= e^{-\frac{1}{2\sigma}\left[\sum_{i=1}^{n} X_i^2 - \mu\sum_{i=1}^{n} X_i + \mu^2\right]} \\ &= e^{-\frac{n}{2\sigma}\left[r_2(X_1,\ldots,X_n) + r_1^2(X_1,\ldots,X_n) + 2\mu r_1(X_1,\ldots,X_n) + \mu^2\right]}. \end{aligned} \tag{4}$$

In this case, the statistics r_1, r_2 are said to be *jointly sufficient*, [DeGroot, 1970] page 158. Note that the set of sufficient statistics is not unique, e.g. in the above example, we can use average of X_i^2 instead of empirical variance.

2.2 Exponential family of distributions

The family of probability distributions having the form:

$$f(x|\theta) = u(x)e^{c(\theta)\cdot d(x) - a(\theta)} \tag{5}$$

where $\theta \in \Theta$ and $c(\theta)$, $d(x)$ and θ are scalars or vectors is called the *exponential* family or sometimes the *Koopman-Darmois family* family. (e.g. [DeGroot, 1986] pages 362, 370 and [DeGroot, 1970] page 161). It follows from factorization criterion (Theorem 1) that $d(x)$ is a sufficient statistics for the given exponential family. Note that $u(x)$ plays a role of a measure on a sample space. Its meaning is demonstrated in Section 2.2.3.

Since $\int_S f(x|\theta)dx = 1$, the value of $a(\theta)$ is completely determined by the value of $c(\theta)$. Thus we can, without loss of generality, use $Image(c)$ instead of Θ as a parameter space. In this way we get the following representation of exponential family, which was used by Schwarz [Schwarz, 1978], with $u(x) \equiv 1$:[1]

$$f(x|\theta) = e^{\theta \cdot y(x) - b(\theta)}. \tag{6}$$

The parameters θ in the above representation are called the *natural* [Kass and Vos, 1997] (or sometimes *canonical* [Barndorff-Nielsen, 1978]) parameters of an exponential family.

We adopt the exponential distribution as presented by Equation 6 and analyze its properties. The moment generating function ([DeGroot, 1970] page 26) of an exponential distribution is:

$$\psi(t) = E(e^{t \cdot y}) = \int e^{t \cdot y} e^{\theta \cdot y - b(\theta)} dy = \int e^{(\theta+t) \cdot y - b(\theta+t)} e^{b(\theta+t) - b(\theta)} dy = e^{b(\theta+t) - b(\theta)} \tag{7}$$

where t is a vector of an appropriate dimension. Therefore

$$\bar{y} = E(y) = \nabla\psi(0) = \nabla b(\theta) \tag{8}$$

and

$$Cov(y) = E\left((y - \bar{y})(y - \bar{y})^T\right) = E(yy^T) - \bar{y}\bar{y}^T = \mathcal{H}(\psi)(0) - \bar{y}\bar{y}^T = \mathcal{H}b(\theta) \tag{9}$$

where $\mathcal{H}$ denotes the Hessian operator, i.e., $\mathcal{H}b(\theta)$ is the matrix of second-order derivatives of $b(\cdot)$ evaluated at θ. Note, that $\mathcal{H}b(\theta)$ is positive definite, unless statistics y are linearly dependent ([DeGroot, 1970] page 25, [Barndorff-Nielsen, 1978] Section 8.1). Here, we suppose that y are linearly independent and Equation 6 is a minimal form for a family of distributions under consideration. For discussion on this topic see [Kass and Vos, 1997] page 15.

Below we give a number of examples of exponential families of distributions.

2.2.1 An Example of Exponential Family - Normal Distribution

Consider a one dimensional normal distribution, [DeGroot, 1970] page 37,

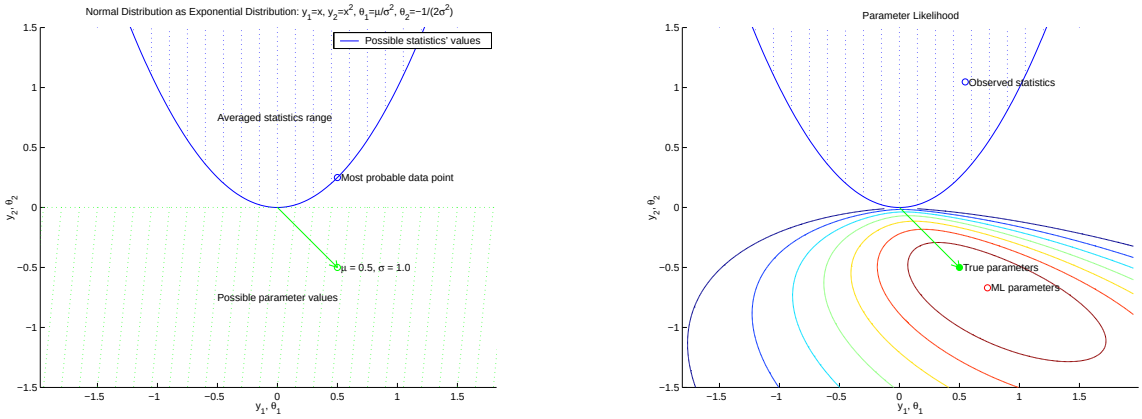
$$f(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]. \tag{10}$$

We can rewrite that in the exponential form:

$$\begin{array}{l} f(x|\mu, \sigma^2) = e^{-b(\theta_1, \theta_2) + \sum_{i=1,2} \theta_i y_i(x)} \\ y_1(x) = x, \quad y_2(x) = x^2, \\ \theta_1 = \frac{\mu}{\sigma^2}, \quad \theta_2 = -\frac{1}{2\sigma^2} \\ b(\theta_1, \theta_2) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\ln\left[-\frac{\pi}{\theta_2}\right] \end{array} \tag{11}$$

In this parameterization, the natural parameter space is $(-\infty, +\infty) \times (-\infty, 0)$ and sufficient statistics y_1 and y_2 get the values from the parabola $y_2 = y_1^2$. Note that it follows from the simple geometric arguments that the most probable data point $\arg\max_y e^{\theta \cdot y - b(\theta)} = \arg\max_y \theta \cdot y$ $(x = \mu)$ corresponds to the point on the parabola where the tangent line is perpendicular to the parameter vector. These concepts are illustrated on Figure 1.

[1] In Schwarz' paper $u(x)$ can be seen as being absorbed into the a priori data probability $P(D)$, since it does depend on θ.



(a) (b)

Figure 1: Representation of normal distribution family as a subfamily of exponential family of distributions. (a) Graph shows the range of natural parameters $(-\infty, +\infty) \times (-\infty, 0)$, the range of possible statistics of each sample (parabola $y_2 = y_1^2$), the range of averaged statistics for sample from normal distribution (parabola interior) and specific parameter vector and most probable data point for these parameters. (b) Graph shows averaged statistics from sampling 100 points from distribution defined by true parameters. It also shows the parameter likelihood isolines and maximum likelihood point given the sampled data.

2.2.2 An Example of Exponential Family - Beta Distribution

Consider a beta distribution which is a one dimensional Dirichlet distribution with parameters $\alpha > 0$ and $\beta > 0$, [DeGroot, 1970] page 40,

$$f(x|\alpha,\beta) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & for\ 0 < x < 1 \\ 0 & otherwise. \end{cases} \tag{12}$$

where $B(\alpha, \beta)$ is the *beta* function [Korn and Korn, 1974] page 743,

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx \tag{13}$$

We can rewrite the beta distribution in the exponential form:

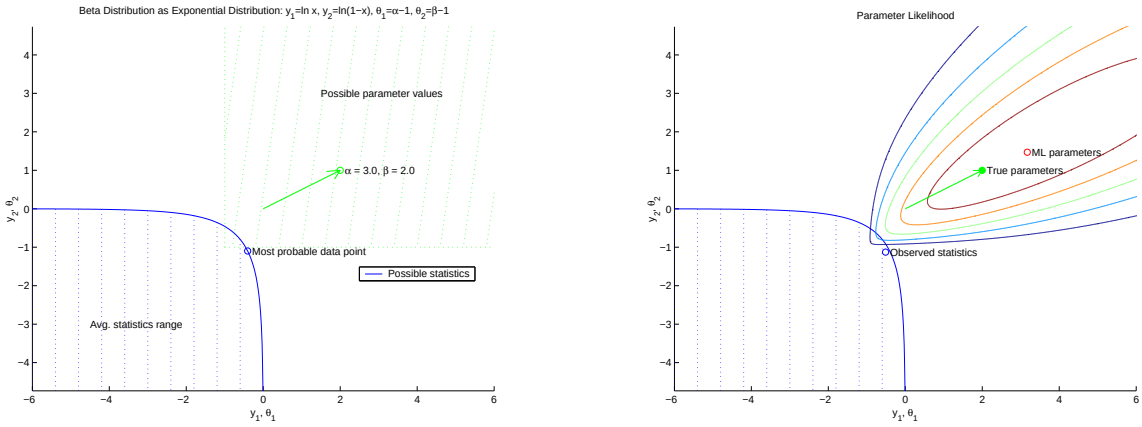
$$\begin{aligned} &f(x|\alpha,\beta) = e^{-b(\theta_1,\theta_2)+\sum_{i=1,2}\theta_i y_i(x)} \\ &y_1(x) = \ln x, \quad y_2(x) = \ln(1-x), \\ &\theta_1 = \alpha - 1, \quad \theta_2 = \beta - 1 \\ &b(\theta_1,\theta_2) = \ln B(\theta_1+1, \theta_2+1) \end{aligned} \tag{14}$$

In this parameterization, the natural parameter space is $(-1, +\infty) \times (-1, +\infty)$ and sufficient statistics y_1 and y_2 get the values from the curve $y_2 = \ln(1 - e^{y_1})$. This concept is illustrated on Figure 2. Once again the most probable data point $x = \frac{\alpha-1}{\alpha+\beta-2}$ corresponds to the point on the curve where the tangent line is perpendicular the parameter vector.

2.2.3 An Example of Exponential Family - Multinomial Distribution

Consider a multinomial distribution, [DeGroot, 1970] page 49, for the distribution of outcomes of n independent trials of discrete random variable with k states. This distribution is defined for $\tilde{\mathbf{x}} = (x_1, \ldots, x_k) \in \mathbb{R}^k$ s.t. $\sum_{i=1}^k x_i = n$ with parameters $\mathbf{p} = (p_1, \ldots, p_k)$, s.t. $\sum_{i=1}^k p_i = 1$ and $\forall i, p_i > 0$:

$$f(\tilde{\mathbf{x}}|n,\mathbf{p}) = \frac{n!}{x_1!\cdots x_k!} p_1^{x_1}\cdots p_k^{x_k} \tag{15}$$



(a) (b)

Figure 2: Representation of the beta distribution family as a subfamily of exponential family of distributions. (a) Graph shows the range of natural parameters $(-1, +\infty) \times (-1, +\infty)$, the range of possible statistics of each sample (curve $y_2 = \ln(1 - e^{y_1})$), the range of averaged statistics for sample from normal distribution (curve interior) and specific parameter vector and most probable data point for these parameters. (b) Graph shows averaged statistics from sampling 100 points from distribution defined by true parameters. It also shows the parameter likelihood isolines and maximum likelihood point given the sampled data.

Note that n is actually a part of a problem definition, like the sample space S. The above distribution can be rewritten into the exponential form

$$\begin{aligned} &f_n(\mathbf{x} = (x_1, x_2, \ldots, x_{k-1})|\eta) = u(\mathbf{x})e^{\eta\cdot\mathbf{x} - nb(\eta)} \\ &u(\mathbf{x}) = \tfrac{n!}{x_1!\cdots x_k!}, \quad x_k = n - \textstyle\sum_{i=1}^{k-1} x_i, \\ &\eta = (\eta_1, \ldots, \eta_{k-1}), \quad \eta_i = \ln \tfrac{p_i}{p_k} \\ &b(\eta) = -\ln p_k = \ln(1 + e^{\eta_1} + e^{\eta_2} + \ldots + e^{\eta_{k-1}}) \end{aligned} \tag{16}$$

Note that $u(\cdot)$ plays the role of measure on the sample space of $S = \{\mathbf{x}|x_i \in \mathbb{Z}, x_i \geq 0, \sum_{i=1}^{k-1} x_i \leq n\}$.

The connection between multinomial distribution to exponential family becomes more apparent if we consider the probabilities of specific series of outcome (taking order into account). It is equivalent to taking $u(\mathbf{x}) \equiv 1$), for $n = 1$ it becomes

$$f_1(y) = e^{\eta\cdot y - b(\eta)} \tag{17}$$

where $y \in S = \{a|a \in \{0, 1\}^{k-1}, |a| \leq 1\}$ and for an (ordered) outcome of series of n experiments, $\mathbf{x} = y_1 + y_2 + \ldots + y_n$, we have:

$$f_n(\mathbf{x}) = \prod_{i=1}^{n} f_1(y_i) = e^{\eta\cdot\mathbf{x} - nb(\eta)} = e^{n(\eta\cdot\rho - b(\eta))} \tag{18}$$

where $\rho = \frac{1}{n}\mathbf{x}$ is the averaged statistics. An illustration of 3-nomial distribution is given on Figure 3.

2.3 L_p norms

We refer the reader to [Doob, 1994] for the basic definitions from the measure theory and the theory of integration, and we state here the basic facts about L_p norms.

Definition. The class L_p (on some measure space $(S, \mathbb{S}, \lambda)$) is the class of measurable functions f for which $|f|^p$ is Lebesgue integrable; and for $p = +\infty$ it is the class of essentially bounded measurable functions, i.e., $\exists c$, s.t. $|f| < c$ almost everywhere.

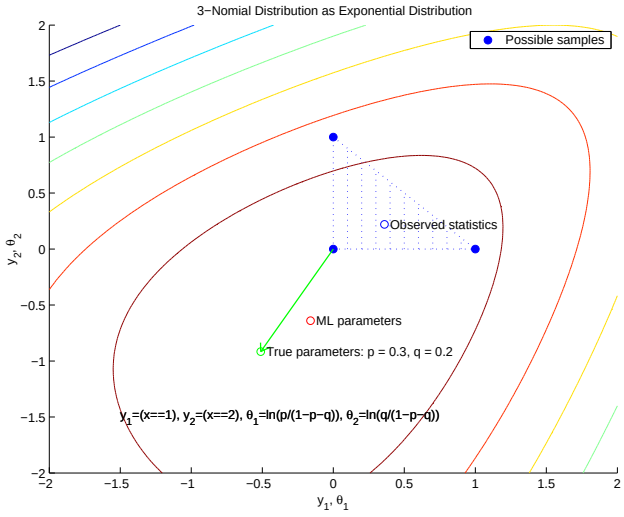


Figure 3: Representation of the 3-nomial distribution family as a subfamily of exponential family of distributions. The range of natural parameter if $\mathbb{R}^2$. Graph shows the possible sample points (black circles), the range of averaged statistics (triangle interior), specific parameter vector and averaged statistics from sampling 100 points from distribution defined by given parameter vector. It also shows the parameter likelihood isolines and maximum likelihood point given the sampled data.

Note that L_p class is linear: a constant multiple of a function in the class is also in the class, and for any positive a, b

$$(a+b)^p \leq 2^p \max(a^p, b^p) \leq 2^p(a^p + b^p) \tag{19}$$

so the sum of two functions in the class is also in the class.

Definition. The notation $\|f\|_p$ stands for L_p *norm of* f for $p < \infty$ and it stands for *essential norm of* f for $p = \infty$. It is

$$\|f\|_p = \begin{cases} \left(\int_S |f|^p d\lambda\right)^{1/p} & p < +\infty \\ ess\sup_S |f| & p = +\infty \end{cases} \tag{20}$$

where $ess\sup f = \sup[\alpha : \lambda\{f \geq \alpha\} \neq 0]$.

Note that for $(S, \mathbb{S}, \lambda)$ being a probability space, i.e. $\lambda(S) = 1$, $\|f\|_p$ is the same as $E^{1/p}\{|f|^p\}$. The well known Jensen's inequality states:

Theorem 2 (Jensen's Inequality) *Let ϕ be a convex function from an interval I of $\mathbb{R}$ into $\mathbb{R}$, and define ϕ at an endpoint of I not in I as the limit of ϕ at the point. Let f be a measurable function from a probability space into I. If f and $\phi(f)$ are integrable, then*

$$\phi\left[E\{f\}\right] \leq E\{\phi(f)\}. \tag{21}$$

If f and $\phi(f)$ are not supposed integrable, but if ϕ and f are lower bounded, Equation 21 remains true.

From Jensen's inequality it follows that, for $1 \leq p < \infty$

$$E^p\left\{|f|\right\} \leq E\left\{|f|^p\right\} \tag{22}$$

which means that $\|f\|_1 \leq \|f\|_p$ for $f \in L_p$. Using substitution $f = |f|^p$ and $p = p'/p$ we get from Equation 22:

$$\|f\|_p \leq \|f\|_{p'} \tag{23}$$

for $f \in L_{p'}$ and $1 \leq p \leq p' \leq +\infty$. The case $p' = +\infty$ is established by direct integration.

Theorem 3 (L_p norm convergence) *Let f be an integrable function on a finite measure space ($\lambda(S) < \infty$), then*

$$\lim_{p\to\infty} \|f\|_p = \operatorname{ess\,sup}_S |f| \tag{24}$$

Proof: Let $c = \operatorname{ess\,sup}_S |f|$ and observe that for any $c' < c$,

$$c' \left[\lambda\{f \geq c'\}\right]^{1/p} \leq \left(\int_S |f|^p d\lambda\right)^{1/p} \leq c[\lambda(S)]^{1/p}. \tag{25}$$

The proof is concluded by observation that first and third terms converge to c' and c as $p \to \infty$. ■

2.4 Multivariate Taylor expansion

Recall that the Taylor series expansion for function of one argument $f(x)$ in x_0, which is analytic in x_0 (i.e. differentiable near x_0), is

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \ldots + \frac{1}{(n-1)!}f^{(n-1)}(x_0)(x - x_0)^{n-1} + R_n(x) \tag{26}$$

The similar formula exists for f being the function of a number of arguments. We cite the results below according to [Korn and Korn, 1974] page 173. Another reference is [Lang, 1993] page 349.

Let Φ be a real-valued function on $\mathbb{R}^n$ and let ∇ denote the gradient operator:

$$\nabla\Phi = \left(\frac{\partial\Phi}{\partial x_1}, \frac{\partial\Phi}{\partial x_2}, \ldots, \frac{\partial\Phi}{\partial x_n}\right)^T \tag{27}$$

Let $(u\cdot\nabla)$ denote a directional derivative operator, i.e., $(u\cdot\nabla)\Phi = (\nabla\Phi)^T u$ and $(u\cdot\nabla)^2\Phi = (u\cdot\nabla)\left[(\nabla\Phi)^T u\right] = u^T\mathcal{H}\Phi u$. $\mathcal{H}$ denotes a Hessian operator (matrix of second-order derivatives). The Taylor expansion for Φ is

$$\Phi(r_0+\Delta r) = \Phi(r_0) + (\Delta r\cdot\nabla)\Phi(r_0) + \frac{1}{2!}(\Delta r\cdot\nabla)^2\Phi(r_0) + \ldots + \frac{1}{(n-1)!}(\Delta r\cdot\nabla)^{n-1}\Phi(r_0) + R_n(r_0+\Delta r) \tag{28}$$

where $R_n(r_0 + \Delta r)$ is a remainder, that can be bounded on some neighborhood $N(r_0)$ of r_0 (analogous to the one dimensional case, [Korn and Korn, 1974] page 145, see also [Lang, 1993] page 350)

$$|R_n(r_0 + \Delta r)| \leq \frac{|\Delta r|^n}{n!} C_{N(r_0)} \tag{29}$$

where $C_{N(r_0)}$ is a constant

$$C_{N(r_0)} = \sup_{r\in N(r_0), \|u\|=1} (u\cdot\nabla)^n\Phi(r) \tag{30}$$

2.5 Linear Algebra, Affine Subspaces and Linear Manifolds

In this section we state and prove a number of basic results from linear algebra:

Theorem 4 *Let A be a real-valued, symmetric and positive semidefinite $n \times n$ matrix, then for any $a \in \mathbb{R}^n$:*

$$\lambda_{\min}\|a\|^2 \leq a^T A a \leq \lambda_{\max}\|a\|^2 \tag{31}$$

and

$$\lambda_{\min}^n \leq \det A \leq \lambda_{\max}^n \tag{32}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ and minimal and maximal eigenvalues of A.

Proof: We prove only one side of inequalities, the other is similar. It is known that for any real-valued symmetric matrix A there exists non-degenerate (full rank) matrix B, such that $B^T = B^{-1}$ and $B^T AB = D$, where D is a diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_n$ ([Usmani, 1987] pages 147, 160, see also [Roman, 1992] page 144). We have for Equation 31:

$$a^T Aa = a^T(BDB^T)a = (B^T a)^T D(B^T a) = \sum_{i=1}^{n} \lambda_i \left[(B^T a)_i\right]^2 \leq \lambda_{\max}(B^T a)^T(B^T a) = \lambda_{\max}\|a\|_2^2. \quad (33)$$

For Equation 32 we are relying on the fact that $\det(AB) = \det(A)\det(B)$ ([Usmani, 1987] page 214), so $\det(A) = 1/\det(A^{-1})$. We have

$$\det(A) = \det(BDB^T) = \det(B)\det(D)\det(B^{-1}) = det(D) = \prod_{i=1}^{n} \lambda_i \leq \lambda_{\max}^n. \quad (34)$$

The second parts of inequalities are similar. ■

We now present the definition of affine subspace ([Roman, 1992] page 43)

Definition. An *affine subspace* of a vector space V is a subset S of V such that $S = b + S' = \{b + s | s \in S'\}$, where $b \in V$ and S' is a subspace of V, i.e., S' is a subset of V closed under linear combinations.

If V is finite dimensional vector space over $\mathbb{R}$ then an affine subspace of V can be represented as $S = \{Av + b | v \in \mathbb{R}^k\}$, where $b \in V$, A is a $\dim(V) \times k$ full rank matrix and $k \leq \dim(V)$ is the dimension of S. The corresponding subspace S' is in this case the k dimensional subspace of V defined by column vectors of A.

Suppose now we have a function $f(v) = y^T v + h(v)$ defined over $\mathbb{R}^K$, we are interested in the form of this function on the affine subspace S:

$$\forall \tilde{v} \in S: \quad \tilde{f}(\tilde{v}) = f(A\tilde{v} + b) = y^T(A\tilde{v} + b) + h(A\tilde{v} + b) = \tilde{y}^T\tilde{v} + \tilde{h}(\tilde{v}) \quad (35)$$

where $\tilde{y} = A^T y$ and $\tilde{h}(\tilde{v}) = y^T b + h(A\tilde{v} + b)$. So the function $\tilde{f}$, which is the function f restricted on the affine subspace S, has the same form as f.

In his original paper, Schwarz uses the term *linear submanifold* of $\mathbb{R}^n$, which essentially denotes on open subset of affine subspace of R^n For the convenience we give a summary of the related differential geometry concepts following [Spivak, 1979a, Spivak, 1979b, Kass and Vos, 1997].

Definition. A mapping is called a *homeomorphism* if it is continuous, one-to-one and has a continuous inverse.

Definition. A metric space M is called a *manifold* if for each $x \in M$ there exist some neighborhood U of x and some integer $n \geq 0$ such that U is homeomorphic to $\mathbb{R}^n$.

Definition. *Linear* manifold is a manifold which is locally isometric to $\mathbb{R}^n$ with its usual metric.

For rigorous definitions of differential geometry concepts see [Berger and Gostiaux, 1988], and for a more engineering-style treating of the same topics see [Murray and Rice, 1993].

2.6 Asymptotic Expansion of Integrals by Laplace Method

Consider the one dimensional integral

$$I(n) = \int_a^b e^{-np(t)} q(t) dt, \quad (36)$$

in which a, b, $p(t)$, and $q(t)$ are independent of the positive parameter n. Suppose we are interested in the asymptotic approximation of $I(n)$ for $n \to \infty$, that is we would like to find $f(n)$ such that $\lim_{n\to\infty} \frac{I(n)}{f(n)} = 1$. This asymptotic relationship is denoted by $I(n) \sim f(n)$, for $n \to \infty$.

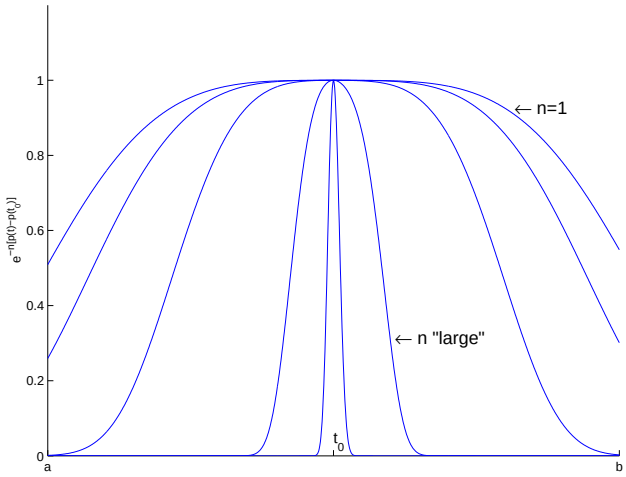


Figure 4: $e^{-n[p(t)-p(t_0)]}$ for various values of n

The following method for approximating $I(n)$ is due to Laplace: The peak value of the factor $e^{-np(t)}$ occurs at $t = t_0$ at which $p(t)$ is a minimum. When n is large, this peak is very sharp and the overwhelming contribution to the integrand comes from the neighborhood of t_0 (Figure 4). We replace $p(t)$ and $q(t)$ by the leading terms in their power (Taylor) expansions around t_0, and extend the integration limits to $\pm\infty$. The resulting integral is explicitly evaluable and yields the required approximation.

Let as illustrate this method, by deriving a first terms in asymptotic approximation of integral given by Equation 36. Note that at minimum of $p(t)$, for $t \to t_0$, $p(t) - p(t_0) \sim \frac{1}{2}p''(t_0)(t-t_0)^2$ and $q(t) \sim q(t_0)$. Also, since $\lim_{t\to t_0}[p(t) - p(t_0) - \frac{1}{2}p''(t_0)(t-t_0)^2] = 0$, then $e^{-n[p(t)-p(t_0)]} \sim e^{-\frac{1}{2}np''(t_0)(t-t_0)^2}$ for $t \to t_0$. Now we have

$$\begin{aligned} I(n) \approx e^{np(t_0)} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}np''(t_0)(t-t_0)^2} q(t_0)dt \quad &= e^{np(t_0)}q(t_0)2\int_0^{+\infty} e^{-u}\frac{1}{\sqrt{2np''(t_0)}}u^{-1/2}du \\ &= e^{np(t_0)}q(t_0)\sqrt{\frac{2}{np''(t_0)}}\Gamma(1/2) = e^{np(t_0)}q(t_0)\sqrt{\frac{2\pi}{np''(t_0)}} \end{aligned} \tag{37}$$

The derivation of the above formula is quite not formal, lacking any proper bounds on the approximation error. We delay the formal treatment of the multidimensional integrals arising in the Bayesian model selection (Equation 1) until Sections 4 and 5 and now we just cite the general results for one dimensional integrals of Laplace type, following [Olver, 1974] page 80. Other good references are [De Bruijn, 1970] Chapter 4 and [Bleistein and Hendelsman, 1975] Sections 5.1 and 8.3.

Without loss of generality it may be supposed that a is finite and the minimum of $p(t)$ on $[a, b]$ occurs at $t = a$; in other cases the integration range can be subdivided at the minima of $p(t)$ and the sign of t reversed where necessary. In addition we suppose:

(i) The minimum of $p(t)$ is approached only at a.

(ii) $p'(t)$ and $q(t)$ are continuous in the neighborhood of a.

(iii) As $t \to a$ from the right: $p(t) - p(a) \sim P \cdot (t-t_0)^\mu$, and $q(t) \sim Q \cdot (t-t_0)^{\lambda-1}$, where P, μ, λ are positive constants, and Q is a real constant.

(iv) $I(n)$ converges absolutely through its range for all sufficiently large n.

and now we can state the following theorem ([Olver, 1974] Theorem 7.1).

Theorem 5 (Laplace Method for One Dimensional Integrals) *With the above conditions:*

$$I(n) \sim \frac{Q}{\mu}\Gamma\left(\frac{\lambda}{\mu}\right)\frac{e^{-np(a)}}{(Pn)^{\lambda/\mu}} \tag{38}$$

The interested readers are referred for the proof of this theorem to [Olver, 1974] page 81 and [Erdélyi, 1956] Section 2.4.

For $p(t)$ being arbitrary function attaining minimum at $t = t_0$ and approximated near t_0 by $p(t) - p(t_0) \sim \frac{1}{2}p''(t_0)(t - t_0)^2$ ($P = \frac{1}{2}p''(t_0)$, $\mu = 2$) and $q(t) \sim q(t_0)$ for $t \to t_0$ ($Q = q(t_0)$, $\lambda = 1$), we have

$$I(n) \sim q(t_0)\Gamma\left(\frac{1}{2}\right)\frac{e^{-np(t_0)}}{\sqrt{\frac{np''(t_0)}{2}}} = q(t_0)\sqrt{\frac{2\pi}{np''(t_0)}}e^{-np(t_0)} \tag{39}$$

which is equivalent to Equation 37 and

$$\ln I(n) \sim -np(t_0) - \frac{1}{2}\ln n + R \tag{40}$$

where $R = \ln q(t_0) + \frac{1}{2}\ln\frac{2\pi}{p''(t_0)}$. As we shall see shortly this is exactly the result of Schwarz for the one dimensional case.

3 The Exact Bayes Procedure

In Bayesian approach to model selection, the model M is chosen according to the maximum posteriori probability given the observed data D:

$$P(M|D) \propto P(M, D) = P(M)P(D|M) = P(M)\int P(D|M, \theta)P(\theta|M)d\theta \tag{41}$$

where θ denotes the model parameters. Instead of maximizing $P(D, M)$ it is convenient to maximize $\ln P(D, M)$. The contribution of [Schwarz, 1978] was the development of an asymptotic expansion of $\ln P(D, M)$. The derivation of this asymptotic expansion is an example of Laplace method for integrals as sketched in the previous section, we show how the results of Schwarz can be extended using the apparatus of Laplace approximation of multivariate integrals ([Bleistein and Hendelsman, 1975], [Hsu, 1948], [Hsu, 1951], [Haughton, 1988]) in Section 5. In this and next sections we present in details the original result of Schwarz.

The asymptotic expansion of $\ln P(D|M)$ is derived under the assumptions stated below. The first two assumptions define the problem settings and the next three assumptions give the sufficient conditions for correctness of presented approximation.

Assumption 1 (Exponential Distribution) *The observations $X_1, \ldots, X_n$ are i.i.d. and come from an exponential distribution:*

$$f(x|\theta) = \exp(\theta \cdot y(x) - b(\theta)). \tag{42}$$

where θ ranges over the natural parameter space Θ, which is a convex subset of $\mathbb{R}^K$.

Assumption 2 (Model linearity) *The models are represented by distinct linear submanifolds of $\mathbb{R}^K$, i.e., any model m_j is a affine subspace of $\mathbb{R}^K$: $\forall\theta \in m_j, \exists\theta^j \in \mathbb{R}^{k_j}$, s.t. $\theta = A_j\theta^j + b_j$, where A_j is a $K \times k_j$ matrix and b_j is a K dimensional vector defined by a model.*

Since m_j is isometric to $\mathbb{R}^{k_j}$ we work with θ^j and $\mathbb{R}^{k_j}$ instead of θ and $m_j \subseteq \Theta$ (see Section 2.5, Equation 35).

Other assumptions taken by Schwarz is the assumption of Bayesian priors, i.e., there is some model that actually describes the observations, and boundedness of conditional densities.

Assumption 3 (Bayesian Prior) *The prior distribution of $\theta \in \Theta$ is of the form:*

$$\mu(\theta) = \sum \alpha_j \mu_j(\theta) \tag{43}$$

where α is the a priori probability of the jth model being the true one, and μ_j is the conditional a priori density of θ given the jth model.

Note that μ (as well as μ_j) stands for a probability density function on Θ.[2]

Assumption 4 (Boundedness) *The conditional a priori distribution of θ given jth model, μ_j, has a k_j dimensional density that is bounded and locally bounded away from zero throughout $m_j \cap \Theta$, i.e., $\exists M$, s.t. $\forall \theta^j \in \mathbb{R}^{k_j}$ s.t. $\theta = A_j\theta^j + b_j \in \Theta \cap m_j$, holds $\mu_j(\theta^j) < M$ (boundedness) and $\forall U \subseteq \mathbb{R}^{k_j}$, U open, $U \subseteq \Theta \cap m_j$, $\exists \epsilon > 0$, s.t. $\forall t \in U, \mu_j(t) > \epsilon$ (local boundedness from zero).*

Assumption 5 (Maximum at Interior Point) [3] *For every m_j, the maximum of $Y \cdot \theta - b(\theta)$ for $\theta \in m_j \cap \Theta$ is achieved at some* interior *point of $m_j \cap \Theta$.*

Assumption 4 ensures that any submanifold of m_j which is of lower dimensionality than m_j will be of measure zero. If, in addition, we assume that any μ_i, μ_j agree on the set $m_i \cap m_j \cap \Theta$, we will also have the 'mutual orthogonality of μ_j', in sense that $m_i \cap m_j \cap \Theta$ has to be of measure zero, since the intersection of two distinct linear manifolds either is one of them (impossible since that will contradict boundedness), or has a lower dimensionality than both.

Note that, unless there is only one model with $k = K$, the parameter prior μ is not absolutely continuous, since it puts positive probability on some lower dimensional submanifolds of Θ, that correspond to the competing models.

Given Assumptions 1, 2 and 3, The Bayes solution for model selection (under fixed penalty for guessing the wrong model) consists of selecting the model that is a posteriori most probable, i.e., choose j that maximizes the posterior model probability (Equation 41). This choice, under Assumptions 1, 2 and 3, corresponds to the maximization of ln of $P(X_1, \ldots, X_n, m_j)$ denoted by $S(Y, n, j)$:

$$S(Y, n, j) = \ln \int_{m_j \cap \Theta} \alpha_j \exp((Y \cdot \theta - b(\theta))n) \mu_j(\theta) d\theta, \tag{44}$$

where Y is the averaged sufficient statistics, i.e. $Y = (1/n) \sum_{i=1}^{n} y(X_i)$.

4 Asymptotics

We are interested in asymptotic expansion of $S(y, n, j)$ as $n \to \infty$.

Proposition 1 *For fixed Y and j, as n tends to ∞,*

$$S(Y, n, j) = n \sup(Y \cdot \theta - b(\theta)) - \frac{1}{2} k_j \ln n + R \tag{45}$$

where the remainder $R = R(Y, n, j)$ is bounded in n for a fixed Y and j.

Note that in case we have a finite number of models, $R(Y, n, j)$ is bounded in n for fixed Y, and bound is independent of j.

Lemma 1 *The proposition holds when $Y \cdot \theta - b(\theta) = A - \lambda \|\theta^j - \theta_0\|^2$ where $\lambda > 0$, θ_0 is a fixed k_j dimensional vector in m_j, and μ_j is Lebesgue measure on $m_j = \mathbb{R}^{k_j}$, i.e. $\mu_j \equiv 1$.*

[2] Schwarz uses μ to denote the probability *measure* on Θ, but since no use is made of the apparatus of the measure theory we simplify the representation and let μ denote the probability density function.

[3] This assumption is not explicitly stated in [Schwarz, 1978], but it is required for the correctness of presented derivations.

Proof: Recall that for normal distribution $(2\pi)^{-k/2}|\Sigma|^{-\frac{1}{2}}\int_{\mathbb{R}^k}\exp\left[-\frac{1}{2}(x-\bar{x})^T\Sigma^{-1}(x-\bar{x})\right]dx = 1$ (e.g. [DeGroot, 1970] page 51), so the explicit evaluation of integral yields

$$\int_{\mathbb{R}^{k_j}} \alpha_j e^{n(A-\lambda\|\theta-\theta_0\|^2)}d\theta = \alpha_j(\pi/n\lambda)^{k_j/2}e^{nA}, \tag{46}$$

and

$$\sup A - \lambda\|\theta-\theta_0\|^2 = A. \tag{47}$$

Therefore

$$S(Y,n,j) = \ln\alpha_j(\pi/n\lambda)^{k_j/2}e^{nA} = nA - \frac{1}{2}k_j\ln n + R \tag{48}$$

establishes the proposition in this case, with $R = \frac{1}{2}k_j\ln\frac{\pi}{\lambda} + \ln\alpha_j$. ■

Lemma 2 *If two bounded positive random variables U and V agree on the set where either exceeds ρ for some $0 < \rho < ess\sup U$, then*

$$\ln E(U^n) - \ln E(V^n) \to 0 \tag{49}$$

as $n \to \infty$.

Proof: We have two bounded positive random variables U and V such that for some ρ, $0 < \rho < \sup U$:

$$U = V \ for\ V > \rho \ or\ U > \rho. \tag{50}$$

Lets first prove the lemma for V that vanishes where $U \le \rho$, i.e., $V = 0$ for $U \le \rho$. In this case $0 \le U^n - V^n \le \rho^n$, and therefore

$$E(V^n) \le E(U^n) \le E(V^n) + \rho^n = E(V^n)\left(1 + \frac{\rho^n}{E(V^n)}\right) \tag{51}$$

it is now sufficient to show that $\ln(1+(\rho^n/E(V^n))) \to 0$ as $n \to 0$. Now $E^{1/n}(V^n) \to ess\sup V$ (Section 2.3 and [Doob, 1994] pages 78, 87). and $ess\sup V = ess\sup U > \rho$ yields that $\rho/(E(V^n))^{1/n}$ a strictly less than 1 (beginning from some n_0), hence $\rho^n/E(V^n)$ tends to zero, and so does $\ln(1+(\rho^n/E(V^n)))$.

Now, for a general V, define $\tilde{V}$ as:

$$\tilde{V} = \begin{cases} V & V > \rho \\ 0 & V \le \rho \end{cases} \tag{52}$$

and similarly $\tilde{U}$. Since $\tilde{V} \le V$ (and $\tilde{U} \le U$) we have

$$-\left(\ln E(V^n) - \ln E(\tilde{U}^n)\right) \le \ln E(U^n) - \ln E(V^n) \le \ln E(U^n) - \ln E(\tilde{V}^n). \tag{53}$$

The right side and the left side of Equation 53 tend to zero as $n \to \infty$ (by the argument from previous paragraph) and thus $\ln[E(U^n) - \ln E(V^n)] \to 0$ as $n \to \infty$. ■

Lemma 3 *For some $0 < \rho < e^A$, where $A = \sup(Y\cdot\theta - b(\theta))$, a vector θ_0, and some positive λ_1 and λ_2, the following holds wherever $e^{Y\cdot\theta-b(\theta)} > \rho$:*

$$A - \lambda\|\theta-\theta_0\|^2 \le Y\cdot\theta - b(\theta) \le A - \lambda\|\theta-\theta_0\|^2. \tag{54}$$

Proof: The matrix of second order derivatives of $b(\theta)$ is the covariance matrix of y and under the assumption of y being linearly independent, $\mathcal{H}b(\theta)$ is positive definite for all θ of interest (see Section 2.2, and [DeGroot, 1970] page 25).

Therefore $Y\cdot\theta - b(\theta)$ is strictly convex, since $\mathcal{H}[Y\cdot\theta - b(\theta)] = -\mathcal{H}b(\theta)$ is negative definite and it attains maximum at θ_0 such that

$$\nabla[Y\cdot\theta_0 - b(\theta_0)] = Y - \nabla b(\theta_0) = 0 \tag{55}$$

i.e., at θ_0 such that $\nabla b(\theta_0) = Y$. Note, that under the Assumption 5, θ_0 is an interior point of $m_j \cap \Theta$.

Consider a Taylor expansion of $\Phi(\theta) = Y \cdot \theta - b(\theta)$ around θ_0 (see Section 2.4 and [Korn and Korn, 1974] page 173):

$$\Phi(\theta_0 + \Delta\theta) = \Phi(\theta_0) + (\Delta\theta \cdot \nabla)\Phi(\theta_0) + \frac{1}{2!}(\Delta\theta \cdot \nabla)^2\Phi(\theta_0) + R_3(\theta_0, \Delta\theta) \tag{56}$$

where $(u \cdot \nabla)$ is a directional derivative operator, i.e., $(u \cdot \nabla)f = (\nabla f)^T u$, and $R_3 = O(\|\Delta\theta\|^3)$ is a remainder, which is bounded by $R_3 < c\|\Delta\theta\|^3$ on some sufficiently small neighborhood of θ_0. Taking the actual derivatives we have

$$\Phi(\theta_0 + \Delta\theta) = Y^T\theta_0 - b(\theta_0) + (Y - \nabla b(\theta_0))^T\Delta\theta - \frac{1}{2}(\Delta\theta)^T\mathcal{H}b(\theta_0)(\Delta\theta) + R_3(\theta_0, \Delta\theta) \tag{57}$$

Since θ_0 is a maximum point, we have $Y - \nabla b(\theta_0) = 0$ and

$$\Phi(\theta_0 + \Delta\theta) = A - \frac{1}{2}(\Delta\theta)^T\mathcal{H}b(\theta_0)(\Delta\theta) + R_3(\theta_0, \Delta\theta) \tag{58}$$

where $A = \sup(Y \cdot \theta - b(\theta))$. Let $\lambda'_1, \ldots, \lambda'_k$ be the eigenvalues of $\mathcal{H}b(\theta_0)$ and let $\lambda_{\min}$ and $\lambda_{\max}$ be the minimal and maximal eigenvalues respectively. We have from Section 2.5

$$\frac{1}{2}\lambda_{\min}\|\Delta\theta\|^2 \leq \frac{1}{2}\Delta\theta^T\mathcal{H}b(\theta_0)\Delta\theta \leq \frac{1}{2}\lambda_{\max}\|\Delta\theta\|^2 \tag{59}$$

Let λ_1, λ_2 be a little bit larger and a little bit smaller than $\frac{1}{2}\lambda_{max}$ and $\frac{1}{2}\lambda_{\min}$ respectively, By strict convexity it is now easy to determine $\rho < e^A$ so that it will bound $e^{Y\cdot\theta - b(\theta)}$ outside the neighborhood $N(\theta_0)$ where $R_3(\theta_0, \Delta\theta)$ is dominated by $(2\lambda_1 - \lambda_{\max})\|\Delta\theta\|^2$ and by $(\lambda_{\min} - 2\lambda_2)\|\Delta\theta\|^2$. I.e, for every $\theta = \theta_0 + \Delta\theta$ from $N(\theta_0)$

$$-\frac{1}{2}(\Delta\theta)^T\mathcal{H}b(\theta_0)(\Delta\theta) + R_3 \leq -\frac{1}{2}\lambda_{\min}\|\Delta\theta\|^2 + cr^3 \leq (-\frac{1}{2}\lambda_{\min} + cr)r^2 \leq -\lambda_2\|\Delta\theta\|^2 \tag{60}$$

where $r = \max_{\theta\in N(\theta_0)}\|\Delta\theta\|$ and λ_2 is such that $\Delta\lambda_2 = \frac{1}{2}\lambda_{\min} - \lambda_2$ is larger than cr. We have

$$A - \lambda_1\|\Delta\theta\|^2 \leq Y \cdot \theta - b(\theta) \leq A - \lambda_2\|\Delta\theta\|^2 \tag{61}$$

where $\theta = \theta_0 + \Delta\theta$ and $2\lambda_1$, $2\lambda_2$ are a little bit larger and a little bit smaller than all the eigenvalues of $\mathcal{H}b(\theta_0)$.

Note, that in order to prove correctness of Proposition 1, the neighborhood defined by ρ should be entirely inside $m_j \cap \Theta$. ■

Proof of the Proposition: For some specific m_j. Let

$$\begin{aligned} U(\theta) &= e^{A-\lambda_1\|\theta-\theta_0\|^2} \\ V(\theta) &= e^{Y\cdot\theta - b(\theta)} \\ W(\theta) &= e^{A-\lambda_2\|\theta-\theta_0\|^2} \end{aligned} \tag{62}$$

Let $N(\theta_0)$ denote the neighborhood of θ_0 where Lemma 3 holds, i.e., where $V(\theta)$ is greater than some $\rho \equiv \rho_V$. Let $\tilde{V}(\theta)$ equal to $V(\theta)$ in $N(\theta_0)$ and zero otherwise. Define $\tilde{U}$ and $\tilde{W}$ similarly. We can bound U, V and W outside that neighborhood by ρ_U, ρ_V and ρ_W and applying Lemma 2 we get

$$\begin{aligned} &S(Y, n, j) - \ln E(\tilde{V}^n) \to 0 \\ &\ln E(U^n) - \ln E(\tilde{U}^n) \to 0 \\ &\ln E(W^n) - \ln E(\tilde{W}^n) \to 0 \end{aligned} \tag{63}$$

From Lemma 3 the following holds on $N(\theta_0)$

$$e^{A-\lambda_1\|\theta-\theta_0\|^2} \leq e^{Y\cdot\theta-b(\theta)} \leq e^{A-\lambda_2\|\theta-\theta_0\|^2} \tag{64}$$

where $A = \sup(Y \cdot \theta - b(\theta))$. Thus for every $\theta \in \mathbb{R}^{k_j}$, $\tilde{U}(\theta) \leq \tilde{V}(\theta) \leq \tilde{W}(\theta)$ and

$$E[\tilde{U}^n(\theta)] \leq E[\tilde{V}^n(\theta)] \leq E[\tilde{W}^n(\theta)] \tag{65}$$

We now evaluate $\ln E[\tilde{U}^n(\theta)]$ and $\ln E[\tilde{W}^n(\theta)]$, i.e., $\ln \int_{N(\theta_0)} e^{n(A-\lambda\|\theta-\theta_0\|^2)}\mu_j(\theta)d\theta$ for $\lambda = \lambda_1, \lambda_2$. Let $c_1 = \min_{N(\theta_0)} \mu_j(\theta)$ and $c_2 = \max_{N(\theta_0)} \mu_j(\theta)$. By application of Lemma 1 we get

$$\ln E[\tilde{W}^n(\theta)] \leq \ln \int_{N_\rho(\theta_0)} e^{n(A-\lambda_2\|\theta-\theta_0\|^2)}\mu_j(\theta)d\theta \leq \ln c_2 + \ln \int_{\mathbb{R}^{k_j}} e^{n(A-\lambda_2\|\theta-\theta_0\|^2)}d\theta = nA - \frac{1}{2}k_j \ln n + R_2 \tag{66}$$

where $R_2 = \frac{1}{2}k_j \ln \frac{\pi}{\lambda_2} + \ln c_2$. Similarly, from Lemma 1 and Lemma 2 we have

$$\begin{aligned} &\ln E[\tilde{U}^n(\theta)] \geq \ln \textstyle\int_{N_\rho(\theta_0)} e^{n(A-\lambda_1\|\theta-\theta_0\|^2)}\mu_j(\theta)d\theta \geq \\ &\quad \ln c_1 + \ln \textstyle\int_{N_\rho(\theta_0)} e^{n(A-\lambda_1\|\theta-\theta_0\|^2)}d\theta \to \ln c_1 + \ln \textstyle\int_{\mathbb{R}^{k_j}} e^{n(A-\lambda_1\|\theta-\theta_0\|^2)}d\theta = nA - \frac{1}{2}k_j \ln n + R_1 \end{aligned} \tag{67}$$

where $R_1 = \frac{1}{2}k_j \ln \frac{\pi}{\lambda_1} + \ln c_1$. Combining Equations 63,65,66 and 67 the proposition is established with R bounded by $\frac{1}{2}k_j \ln \frac{\pi}{\lambda_2} + \ln c_2 + \ln \alpha_j$. ■

Note, that as $n \to \infty$, $\rho \to A$ (smaller $N(\theta_0)$'s) and $\lambda_1 \to \frac{1}{2}\lambda_{\max}$, $\lambda_2 \to \frac{1}{2}\lambda_{\min}$ (maximum and minimum eigenvalues of $\mathcal{H}b(\theta_0)$), the remainder R is approximation of $S(Y, n, j)$ is asymptotically bounded by:

$$\frac{1}{2}k_j \ln \frac{2\pi}{\lambda_{\max}} + \ln \mu(\theta_0) + \ln \alpha_j \leq R \leq \frac{1}{2}k_j \ln \frac{2\pi}{\lambda_{\min}} + \ln \mu(\theta_0) + \ln \alpha_j \tag{68}$$

so the error of approximation of $S(Y, n, j)$ dependents on the minimal eigenvalue of $\mathcal{H}b(\theta_0)$ (e.g. how sharp is the peak at θ_0) and the "freedom" in R depends on the condition number of $\mathcal{H}b(\theta_0)$ (a ratio between maximal and minimal eigenvalues). Note that the above result is a particular application of Laplace method for integrals (Section 2.6). We provide a more accurate estimate for $S(Y, n, j)$ in the next section.

5 Discussion on High-Order Laplace Approximations and Extensions of BIC

Consider again the expression of $P(D|M)$ for exponential models (Equation 44):

$$P(D|M) = I(Y, n, j) = \int_{m_j \cap \Theta} e^{n(Y\cdot\theta - b(\theta))}\mu_j(\theta)d\theta. \tag{69}$$

In the previous section the asymptotic expansion of $I(Y, n, j)$ for a fixed Y was found, i.e.,

$$I(Y, n, j) = e^{\mathrm{BIC}(Y,n,j)}[1 + O(1)], \tag{70}$$

where $\mathrm{BIC}(Y, n, j) = n \sup[Y \cdot \theta - b(\theta)] - \frac{k}{2} \ln n$. Thus $S(Y, n, j) \triangleq \ln I(Y, n, j) = BIC(Y, n, j) + O(1)$. In this section address the following questions:

- Consistency of BIC model selection.
- Asymptotic expansion of $I(Y, n, j)$ for m_j being smooth submanifold of $\mathbb{R}^K$ of dimension k_j.
- Asymptotic expansion of $I(Y, n, j)$ for fixed Y with relative error that tends to zero as $n \to \infty$.
- Asymptotic expansion of $I(Y, n, j)$ for $Y \to Y_0$ as $n \to \infty$.

These questions were addressed in [Haughton, 1988]. We will address them very briefly below, deriving the answers using the machinery of Laplace approximation of multivariate integrals ([Bleistein and Hendelsman, 1975] Section 8.3, [Hsu, 1948], [Hsu, 1951]).

The consistency of BIC model selection: It was shown by [Haughton, 1988] that model selection based on the BIC score is consistent, i.e., if the true parameter θ_0 belongs to model m_1 and does not belong to $m_1 \cap m_2$ then model m_1 will be chosen with probability converging to one as $n \to \infty$. In case $\theta_0 \in m_1 \cap m_2$ the model of lower dimension will be chosen.

Asymptotic expansion of $I(Y, n, j)$ for curved exponential models: This question is rather trivial, since the change of coordinates to the coordinates of smooth manifold of the curved model results in the multiplication of exponential density by Jacobian of transformation, which effectively plays the role of changing the a priori density μ_j. I.e., let ds denote the differential element of surface $m_j \cap \Theta$ and let $x \in \mathbb{R}^{k_j} \cap S$ denote the k_j dimensional parameterization of $m_j \cap \Theta$, we have

$$I(Y, n, j) = \int_{m_j \cap \Theta} e^{n(Y \cdot \theta - b(\theta))} \mu_j(\theta) ds = \int_{\mathbb{R}^{k_j} \cap S} e^{n(Y \cdot \theta(x) - b(\theta(x)))} \mu_j(\theta(x)) \left(\frac{\partial \theta}{\partial x} \right) dx \tag{71}$$

Asymptotic expansion of $I(Y, n, j)$ for fixed Y with higher order term than BIC: We use the results of Laplace approximation of multivariate integrals [Bleistein and Hendelsman, 1975] Section 8.3. Let $\phi(x) = Y \cdot \theta(x) - b(\theta(x))$ denote the function in the exponent for some curved exponential model m_j, and $g(x) = \mu_j(\theta(x)) \left(\frac{\partial \theta}{\partial x} \right)$. We have (omitting the exponentially small error terms that result from finite domain of integration)

$$I(Y, n, j) = \int_{\mathbb{R}^{k_j}} e^{n\phi(x)} g(x) dx = e^{n\phi(x_0)} \left[\frac{g(x_0)}{|\det(\phi_{x_i x_j}(x_0))|} \left(\frac{2\pi}{n} \right)^{n/2} + \frac{1}{n} \int_{\mathbb{R}^{k_j}} G_1(\xi) e^{-\frac{n}{2} \xi \cdot \xi} d\xi \right], \tag{72}$$

where $G_1(\xi)$ is some particular function. From the above formula we can see that BIC together with the next terms $\ln g(x_0) - \frac{1}{2} \ln |\det(\phi_{x_i x_j}(x_0))|$ has an absolute error of approximation $O(\frac{1}{n})$.

Asymptotic expansion of $I(Y, n, j)$ for $Y \to Y_0$ as $n \to \infty$: The asymptotic expansion of $I(Y, n, j)$ for $Y \to Y_0$ is essentially the same as the one given by Equation 72, it is reasonable to suppose that $G_1(\xi)$ does not change a lot for Y very close to Y_0 and thus the second uniformly integral is bounded for all such Y. This in turn implies that Equation 72 is correct for approximation for $I(Y, n, j)$ for $Y \to Y_0$ with relative error of $O(\frac{1}{n})$.

It is interesting to note that [Haughton, 1988] arrives to the same leading terms of asymptotic expansion for $I(Y, n, j)$ for $Y \to Y_0$ as Equation 72, but the error bound provided is only $O(\frac{1}{\sqrt{n}})$. We are convinced that the true error bound is $O(\frac{1}{n})$ and it will be interesting to prove it formally.

6 Acknowledgments

Dmitry Rusakov would like to thank Mark Zlochin for helpful discussions.

References

[Barndorff-Nielsen, 1978] Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families In Statistical Theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons.

[Berger and Gostiaux, 1988] Berger, M. and Gostiaux, B. (1988). *Differential Geometry: Manifolds, Curves, and Surfaces*. Number 115 in Graduate Texts in Mathematics. Springer-Verlag.

[Bleistein and Hendelsman, 1975] Bleistein, N. and Hendelsman, R. A. (1975). *Asymptotic Expansions of Integrals*. Holt, Rinehart and Winston.

[Cooper and Herskovits, 1992] Cooper, G. F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347.

[De Bruijn, 1970] De Bruijn, N. G. (1970). *Asymptotic Methods in Analysis*, volume 4 of *A Series of Monographs on Pure and Applied Mathematics*. North-Holland Publishing Company, 3rd edition.

[DeGroot, 1970] DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill Book Company.

[DeGroot, 1986] DeGroot, M. H. (1986). *Probability and Statistics*. Addison-Wesley, 2nd edition.

[Doob, 1994] Doob, J. L. (1994). *Measure Theory*. Number 143 in Graduate Texts in Mathematics. Springer-Verlag.

[Erdélyi, 1956] Erdélyi, A. (1956). *Asymptotic Expansions*. Dover, New York.

[Geiger et al., 2001] Geiger, D., Heckerman, D., King, H., and Meek, C. (2001). Stratified exponential families: Graphical models and model selection. *Annals of Statistics*, 29(2):505–529.

[Haughton, 1988] Haughton, D. (1988). On the choice of a model to fit data from an exponential family. *Annals of Statistics*, 16(1):342–355.

[Heckerman, 1995] Heckerman, D. (1995). A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Microsoft, One Microsoft Way, Redmond, WA 98052, USA.

[Heckerman et al., 1995] Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243.

[Hsu, 1948] Hsu, L. C. (1948). A theorem on the asymptotic behavior of a multiple integral. *Duke Mathematical Journal*, pages 623–632.

[Hsu, 1951] Hsu, L. C. (1951). On the asymptotic evaluation of a class of multiple integrals involving a parameter. *American Journal of Mathematics*, 73(3):625–634.

[Kass and Vos, 1997] Kass, R. E. and Vos, P. W. (1997). *Geometrical Foundations of Asymptotic Inference*. Wiley Series in Probability and Statistics. John Wiley & Sons.

[Korn and Korn, 1974] Korn, G. A. and Korn, T. M. (1974). *"Spravochnik po Mathematike" a translation of "Mathematical Handbook"*. Nauka.

[Lang, 1993] Lang, S. (1993). *Real and Functional Analysis*. Number 142 in Graduate Texts in Mathematics. Springer-Verlag, 3rd edition.

[Murray and Rice, 1993] Murray, M. K. and Rice, J. W. (1993). *Differential Geometry and Statistics*. Number 48 in Monographs and Statistics and Applied Probability. Chapman and Hall.

[Olver, 1974] Olver, F. W. J. (1974). *Asymptotics and Special Functions*. Computer Science and Applied Mathematics: A Series of Monographs and Textbooks. Academic Press.

[Roman, 1992] Roman, S. (1992). *Advanced Linear Algebra*. Number 135 in Graduate texts in mathematics. Springer-Verlag.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

[Spivak, 1979a] Spivak, M. (1979a). *A Comprehensive Introduction to Differential Geometry*, volume 1. Publish or Perish, Inc., 2nd edition.

[Spivak, 1979b] Spivak, M. (1979b). *A Comprehensive Introduction to Differential Geometry*, volume 2. Publish or Perish, Inc., 2nd edition.

[Usmani, 1987] Usmani, R. A. (1987). *Applied Linear Algebra*. Number 105 in Monographs and Textbooks in Pure and Applied Mathematics. Marcel Dekker.

[Watanabe, 2000] Watanabe, S. (2000). Algebraic analysis for non-regular learning machines. In Solla, S. A., Leen, T. K., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems 12: Proceedings of the 1999 conference*, pages 356–362. MIT Press.