

Algebraic Geometry and Singular Statistics

Algebraic Statistics and Computational Biology
Workshop at Clay Mathematics Institute
November 12-14, 2005

Sumio Watanabe

Tokyo Institute of Technology





Contents

1. Why algebraic geometry ?
2. Mathematical Problems
3. Resolution of Singularities
4. Empirical process \rightarrow ML and MAP
5. Zeta function \rightarrow Bayes
6. Application to Singular Statistics



1. Why algebraic geometry ?

Statistical model

$$p(x|w) \quad x \in \mathbb{R}^N \quad w \in \mathbb{R}^d$$

Identifiable:

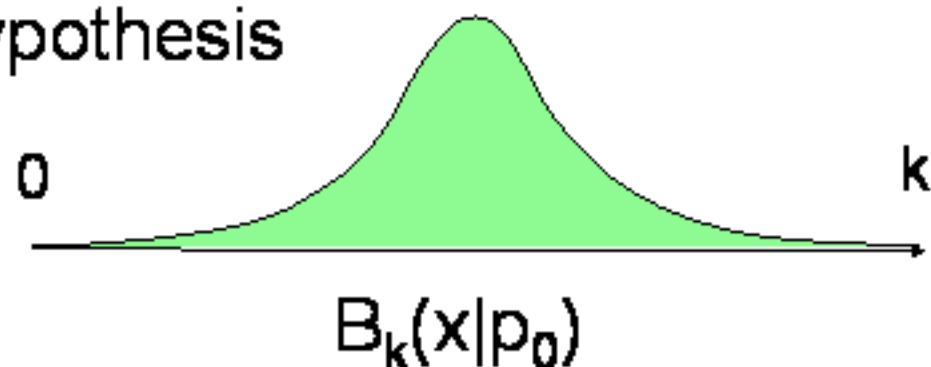
$$p(x|w_1) = p(x|w_2) \quad (\forall x) \Rightarrow w_1 = w_2$$

*If $p(x|w)$ contains hidden variables
or layered structure, then it is nonidentifiable.*

Example.1 : Mixtures of Binomial Dist.

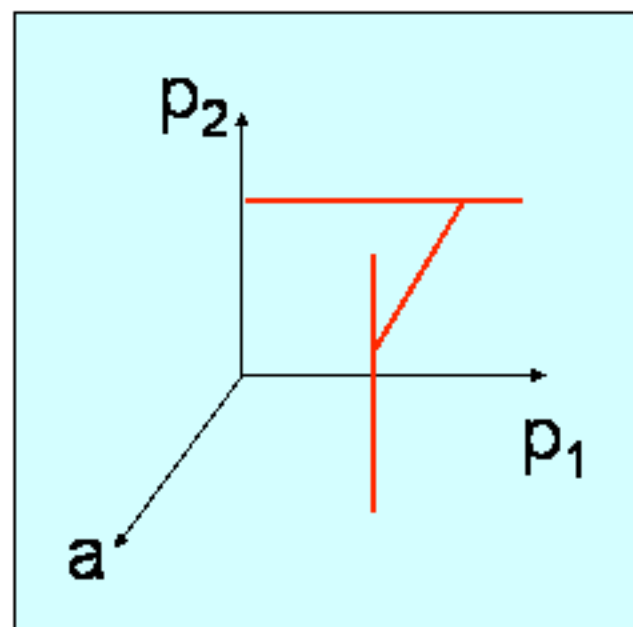
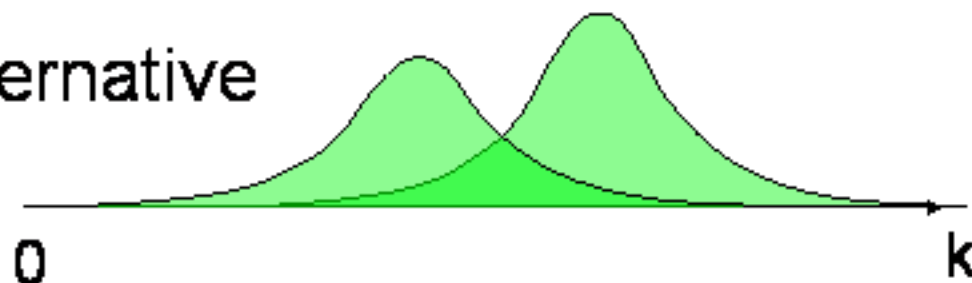


Null
Hypothesis



$$a B_k(x|p_1) + (1-a) B_k(x|p_2)$$

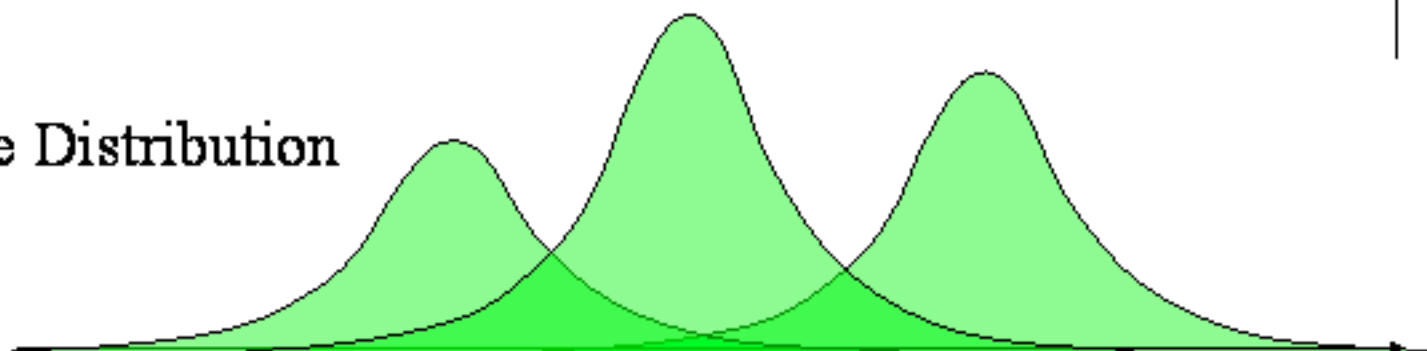
Alternative





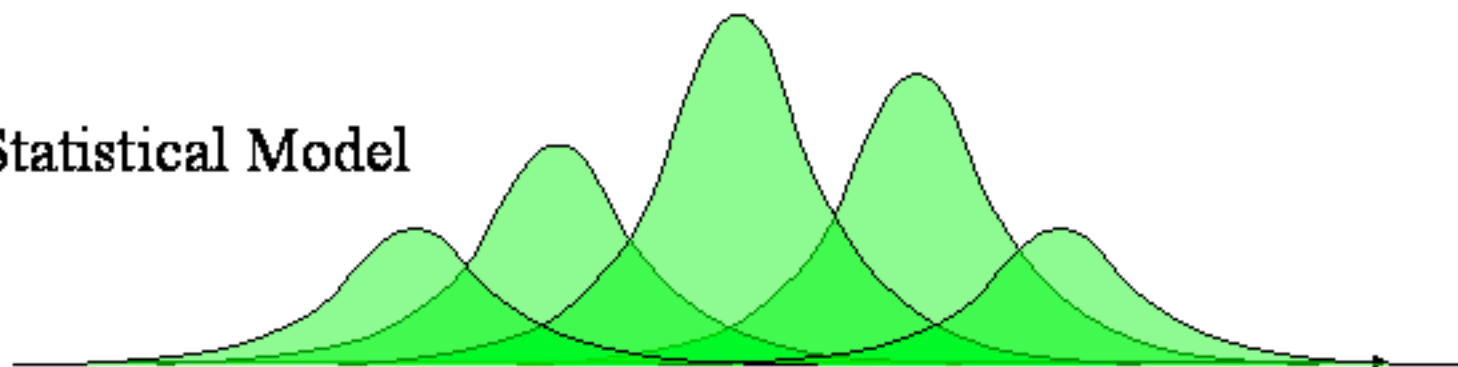
Example.2 : Normal Mixtures

True Distribution



$$p(x|w) = \sum a_h \exp(-||x-b_h||^2)$$

Statistical Model

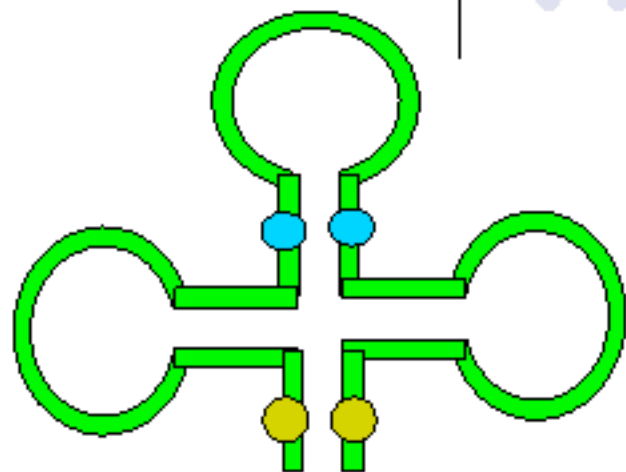


Example.3: Prob. Context-Free Grammar



a b a c a b b c c a ... a c a

a b { a c a (b b c c a) ... a } c a



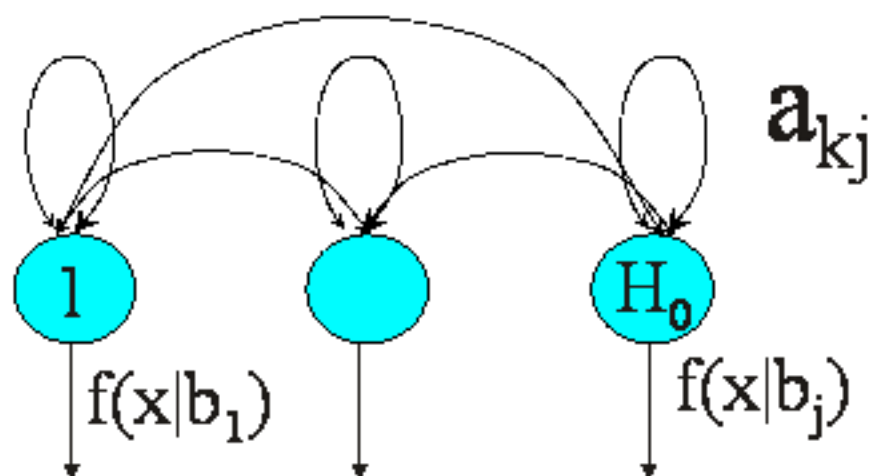
$A \rightarrow AA$
 $A \rightarrow a \mid b$

$A \rightarrow AA \mid AB \mid BA \mid BB$
 $B \rightarrow AA \mid AB \mid BA \mid BB$
 $A \rightarrow a \mid b$
 $B \rightarrow a \mid b$

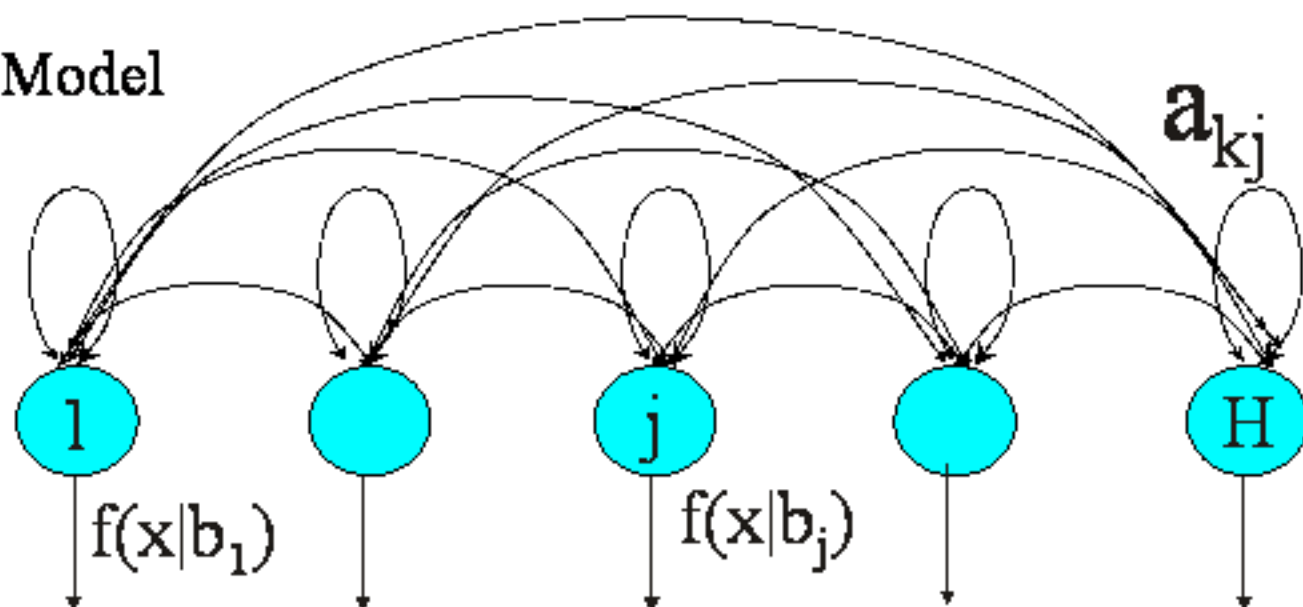
Example.4 : Hidden Markov Models



True distribution



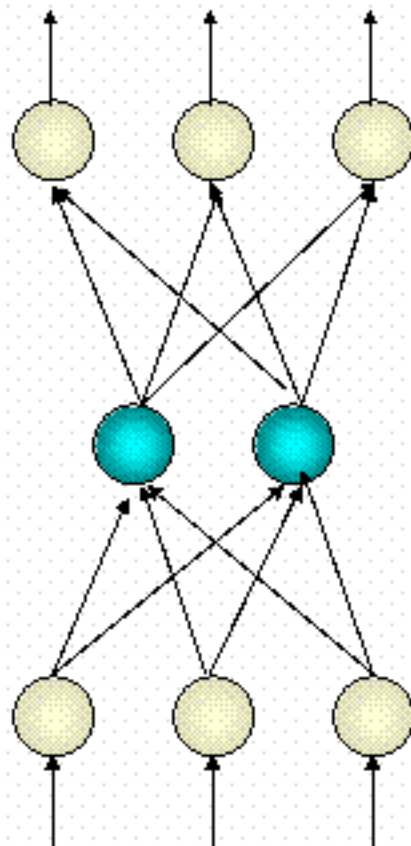
Statistical Model



Example.5 : Neural Networks

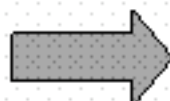


$$y = (y_1, y_2, \dots, y_N)$$

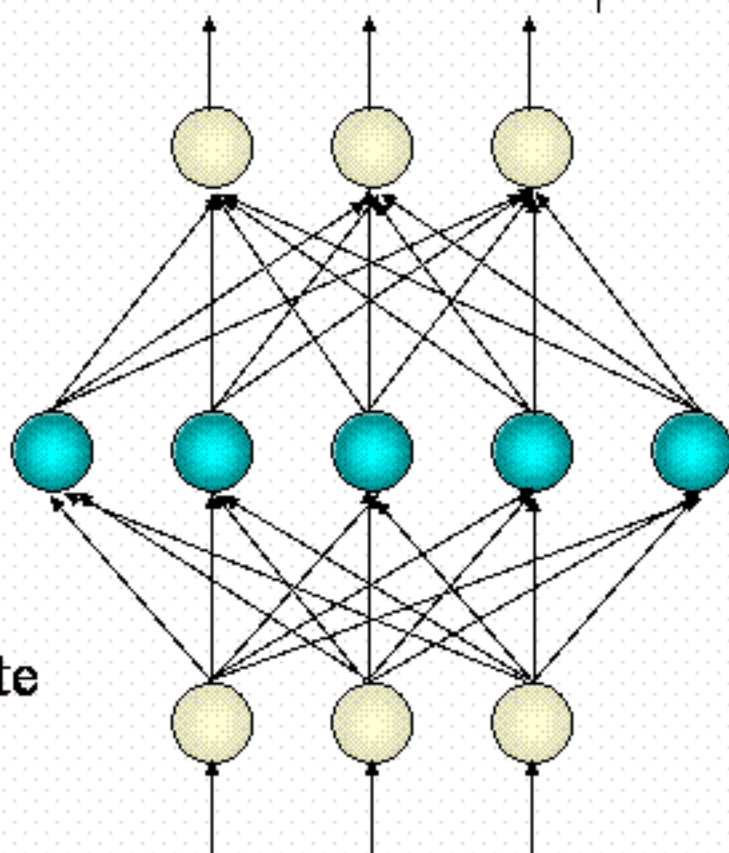


$$x = (x_1, x_2, \dots, x_M)$$

samples



$$y = (y_1, y_2, \dots, y_N)$$



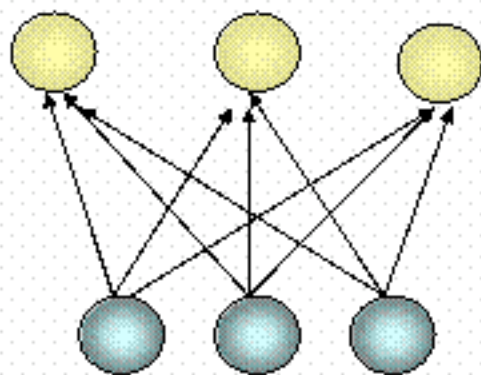
estimate

$$x = (x_1, x_2, \dots, x_M)$$



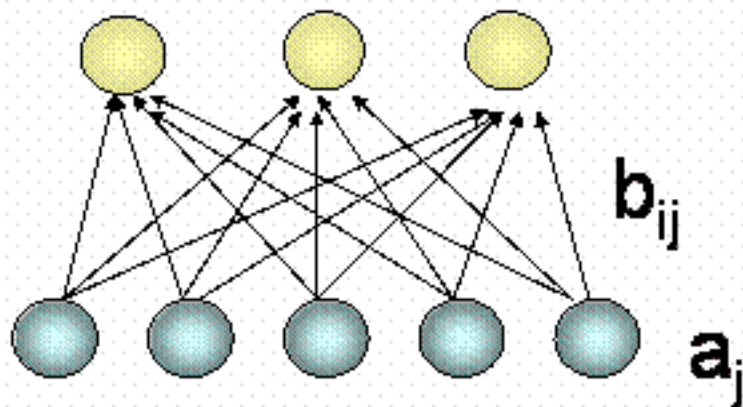
Example.6 : Bayesian Networks

Observables



Hidden variables

Observables



Hidden variables





Algebraic geometry is needed

Equivalence Relation :

$$p(x|w_1) = p(x|w_2) \quad (\forall x) \Leftrightarrow w_1 \sim w_2$$

\mathbb{R}^d / \sim is not a manifold

*What is the appropriate coordinate
for singular statistical models ?*



2. Mathematical Problems

True distribution $p(x|w_0) \rightarrow X_1, X_2, \dots, X_n$

Statistical model $p(x|w)$

Log likelihood ratio function $f(x, w) = \log \frac{p(x|w_0)}{p(x|w)}$

Log likelihood ratio process $L(w) = \sum_{i=1}^n f(X_i, w)$

Mathematical Problem (1)



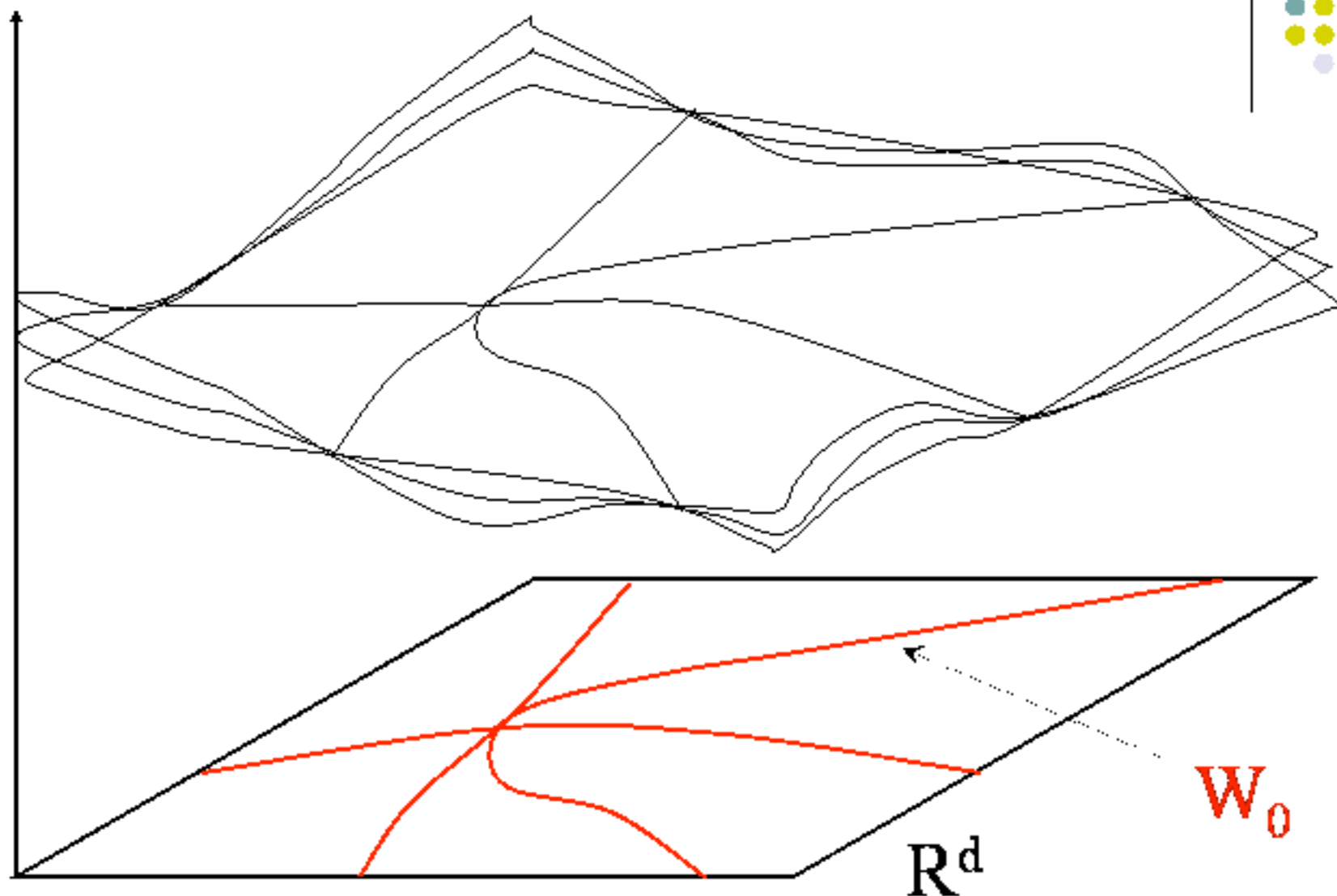
Relative Entropy:

$$\begin{aligned} K(w) &\equiv E_x[f(X, w)] \\ &= \int p(x|w_0) \log \frac{p(x|w_0)}{p(x|w)} dx \geq 0 \end{aligned}$$

$$W_0 \equiv \{ w \in W ; K(w)=0 \}$$

: an analytic set with singularities

$L(w)$: log likelihood ratio



Mathematical Problem (1)



$W_0 = \{w; K(w) = 0\}$ is an analytic set with singularities.

Clarify the behavior of the random process on W ,

$$L(w) = \sum_{i=1}^n f(X_i, w)$$

→ **ML, MAP, and Hypothesis Testing**

Mathematical Problem (2)



$K(w) = 0$ is an analytic set with singularities.

Clarify the behavior of the random variable,

$$F = -\log \int \exp(-L(w)) \underbrace{\phi(w)}_{\text{Prob. Dist. on } W} dw$$

$$L(w) = \sum_{i=1}^n f(X_i, w)$$

*Prob. Dist. on W
Called prior*

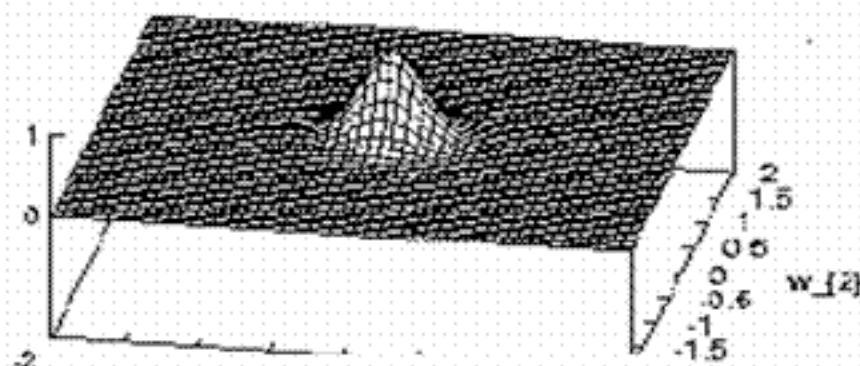
F = Stochastic Complexity \rightarrow Bayes estimation

Essential Difference between regular and singular models

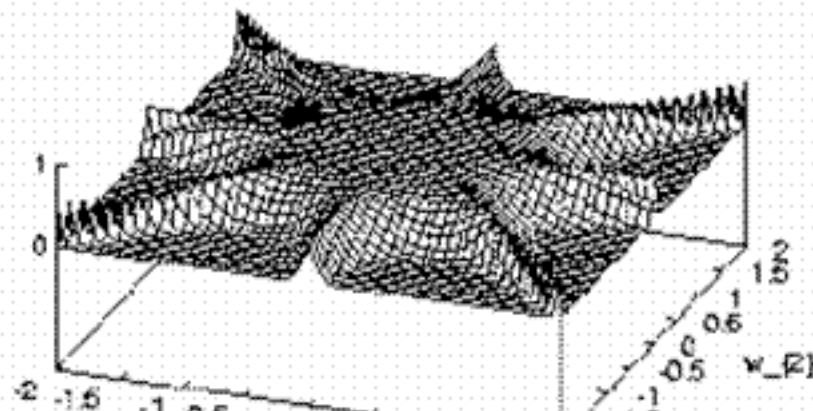


Likelihood function $\exp(-L(w))$

Bayes A Posteriori Dist. $\exp(-L(w)) \phi(w)$



Regular Models
Fisher, Cramer, 1930



Singular Models



Mathematical Conditions

$$f(x, w) : w \in W \subset \mathbb{R}^d$$

W : compact

Analytic function from W to $L^2(p(x|w_0)dx)$

$$\int |f(x, w)|^2 p(x|w_0)dx < \infty$$



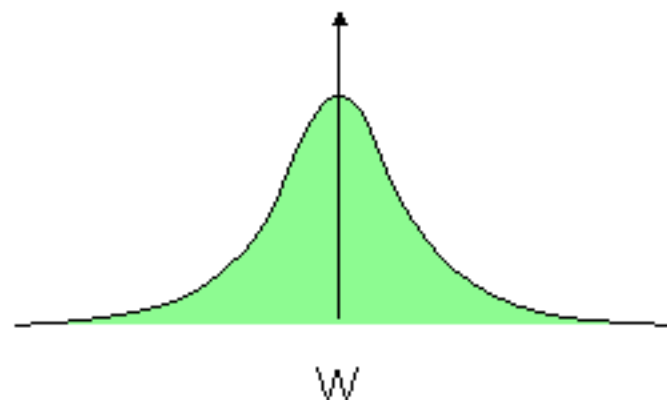
Example: A regular case

$$p(x|w) = (2\pi)^{-1/2} \exp(-1/2(x-w)^2)$$

$$w_0 = 0$$

$$f(x, w) = w^2/2 - wx$$

$$K(w) = w^2/2$$



$$\begin{aligned} L(w) &= \sum_{i=1}^n f(X_i, w) \\ &= nK(w) + (nK(w))^{1/2} \left(\left(\frac{2}{n} \right)^{1/2} \sum_{i=1}^n X_i \right) \end{aligned}$$

How can we generalize this to singular statistical models ?



Singular Statistical Models :

$$\psi_n(w) = \frac{1}{(nK(w))^{1/2}} \sum_{i=1}^n (f(X_i, w) - K(w))$$

$$L(w) = nK(w) + (nK(w))^{1/2} \psi_n(w)$$

Two Problems:

- (1) $K(w)=0$ contains singularities.
- (2) $\psi_n(w)$ is not well-defined near singularities.

3. Resolution of Singularities (1964, Hironaka)

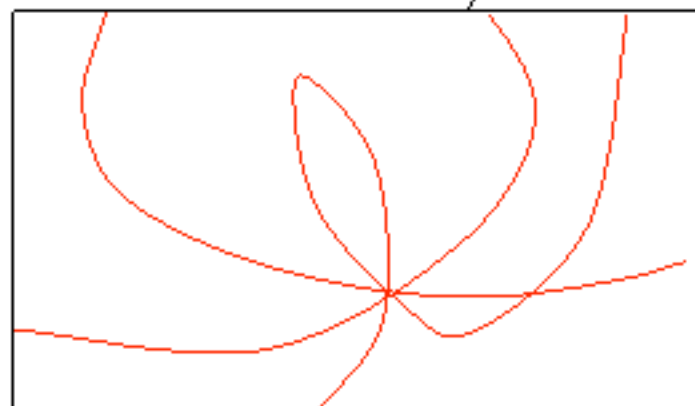


0 $\xrightarrow{\hspace{10em}}$ ∞

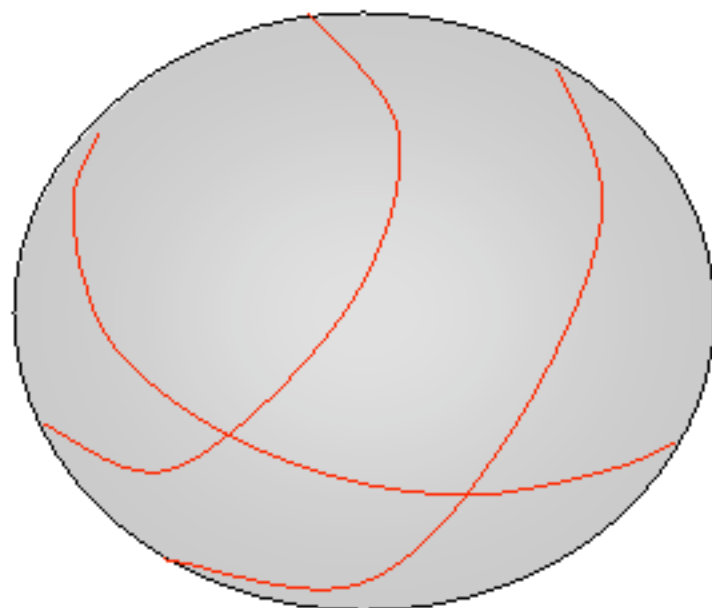
Analytic
nonnegative
 $K(w)$

In every local coordinate,
Normal crossing

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



proper
analytic
 $w = g(u)$



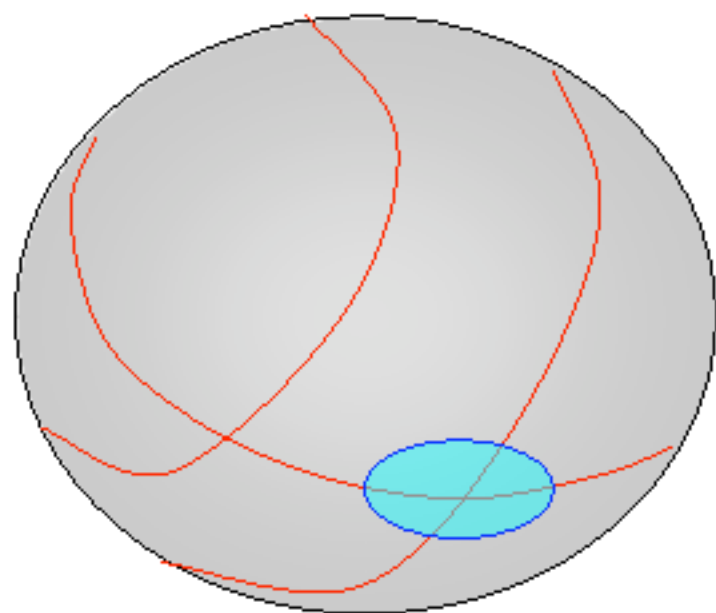
W compact \subseteq Euclidean Space

U compact \subseteq Manifold



Decomposition of Log Likelihood ratio

In every local coordinate,



$U \subseteq \text{Manifold}$

$$K(g(u)) = \prod_{i=1}^d u_i^{2k_i}$$

Jacobian

$$|g(u)'| = b(u) \prod_{i=1}^d |u_i^{h_i}|$$

$$f(x, g(u)) = a(x, u) \prod_{i=1}^d u_i^{k_i}$$



4. Empirical Process

$$\psi_n(g(u)) = \frac{1}{n^{1/2}} \sum_{i=1}^n (a(X_i, u) - \Pi u_i^{k_i})$$

made to be well-defined by selecting a branch.

C(U): the set of continuous functions on compact **U**
separable and complete metric space with

$$\|f\| = \max_{u \in U} |f(u)|$$



$(C(U), B, \psi_n)$ Probability Space

$\psi_n \rightarrow \psi$: convergence in law

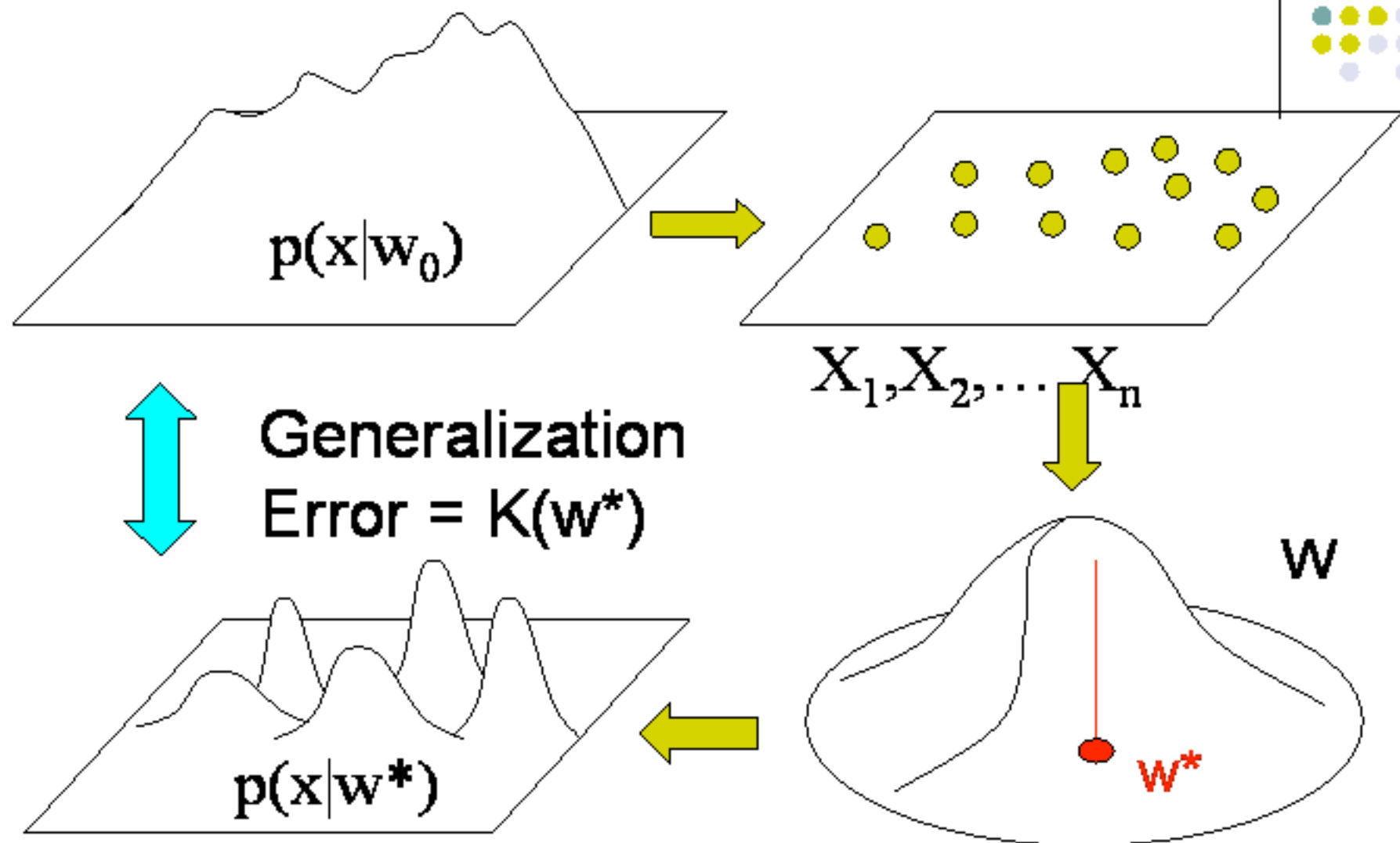
i.e. $E[F(\psi_n)] \rightarrow E[F(\psi)]$ for F : bdd & cont

ψ : tight gaussian process on U

i.e. $\forall \varepsilon > 0, \exists K$ compact $\subseteq C(U)$ s.t. $P(K) > 1 - \varepsilon$.

$$L(g(u)) \rightarrow n \prod u_i^{2k_i} + n^{1/2} \prod u_i^{k_i} \psi(u)$$

ML and MAP estimation



Application to ML and MAP



Log Likelihood ratio is minimized at $u = u^*$.

$$L(g(u)) = n \left\{ \prod u_i^{k_i} - n^{-1/2} \psi(u) / 2 \right\}^2 - |\psi(u)|^2 / 4$$

Generalization Error

$$K(g(u^*)) = \prod u_i^{*2k_i} \longrightarrow (1/4n) \max_{K(g(u))=0} |\psi(u)|^2$$

Empirical Error

$$(1/n)L(g(u^*)) \longrightarrow - (1/4n) \max_{K(g(u))=0} |\psi(u)|^2$$



Theorem 1.

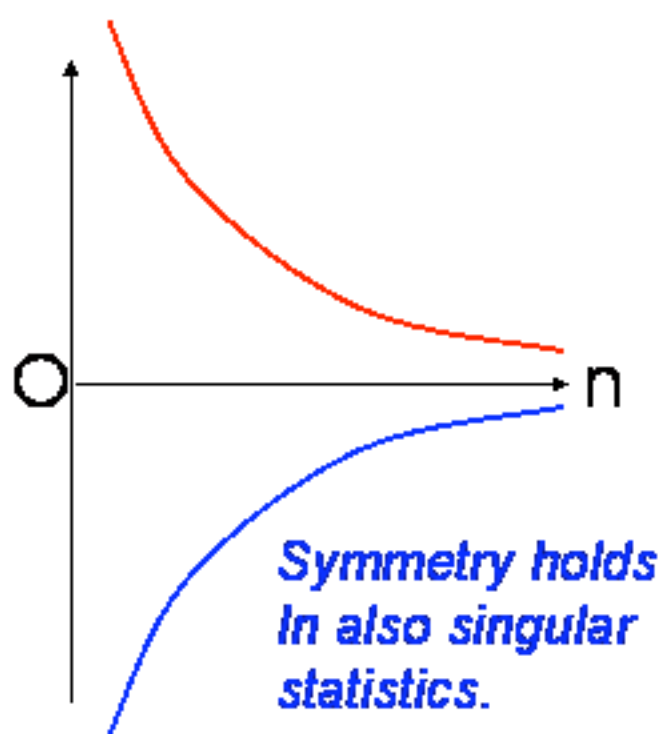
In the ML and MAP estimations,
the generalization and the empirical error
are respectively given by

Generalization Error

$$= (1/4n) \max_{K(g(u))=0} |\psi(u)|^2$$

Empirical Error

$$= - (1/4n) \max_{K(g(u))=0} |\psi(u)|^2$$



5. Zeta, State density, and partition functions



zeta function

$$\zeta(z) = \int K(w)^z \phi(w) dw \xrightarrow{\text{Resolution of singularities}} \frac{C_1}{(z + \lambda)^m}$$

State density function

$$v(t) = \int \delta(t - K(w)) \phi(w) dw = C_2 t^{\lambda-1} (-\log t)^{m-1}$$

$t \rightarrow 0$

Partition function

$$Z(n) = \int \exp(-nK(w)) \phi(w) dw = \frac{C_3 (\log n)^{m-1}}{n^\lambda}$$

$n \rightarrow \infty$

This structure was found by Gel'fand, Atiyah, Bernstein, Sato, and Kashiwara
1954, 1970, 1972, 1974, 1978

Desingularization \rightarrow Analytic continuation

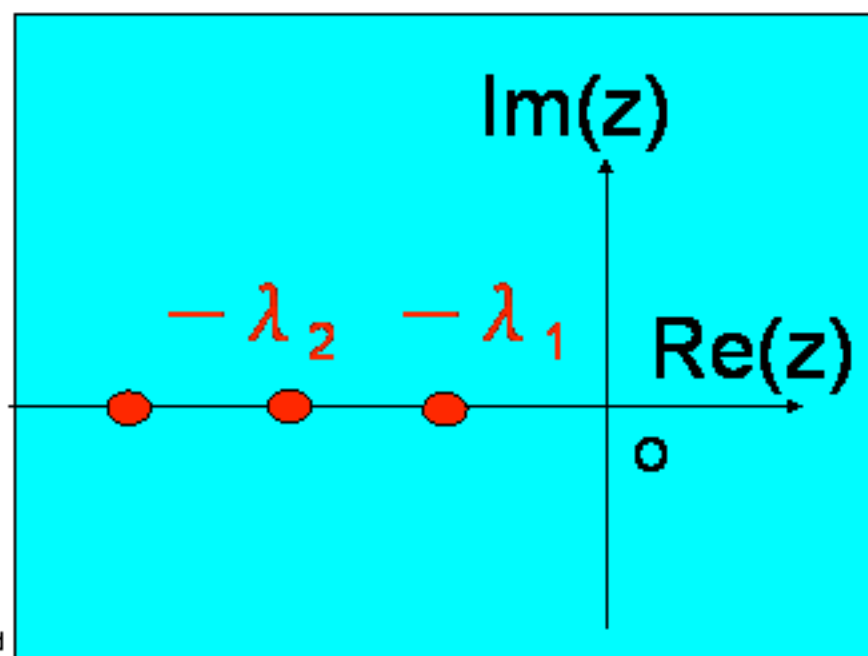


$$\xi(z) = \sum \int K(g(u))^z \phi(g(u)) |g'(u)| du$$

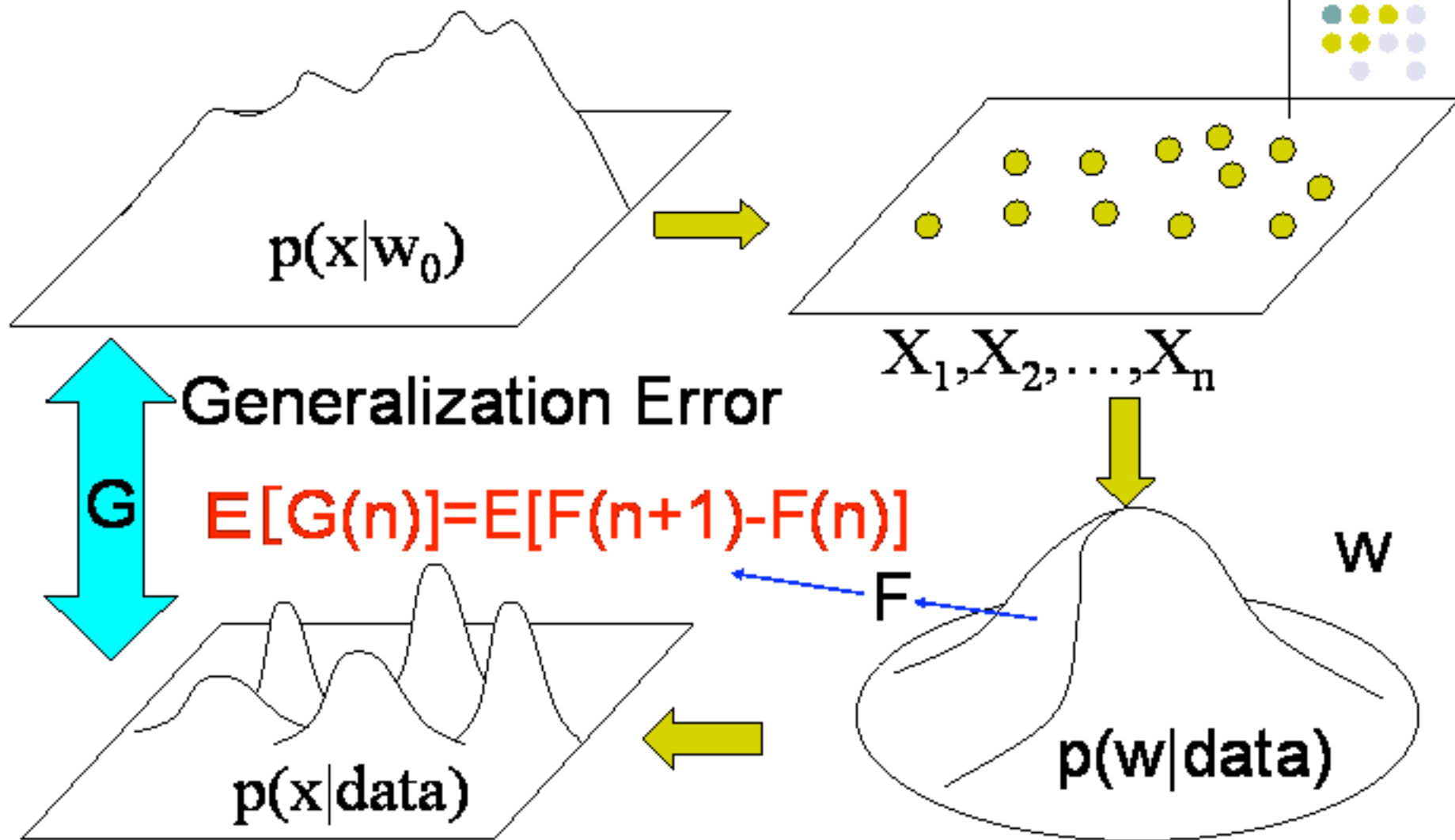
$$= \sum \int \prod u_j^{k_j} z^{+h_j} \phi(g(u)) du$$

$$= \frac{C_1}{(z + \lambda_1)^{m_1}} + \frac{C_2}{(z + \lambda_2)^{m_2}} + \dots$$

$\xi(z)$, which is a holomorphic function in $\text{Re}(z) > 0$, can be analytically continued to entire complex plane as a meromorphic function.



Bayes estimation





Application to Bayes Estimation

Log marginal likelihood = Stochastic complexity

$$F = -\log \int \exp(-L(w)) \phi(w) dw$$
$$= -\log \sum \int \exp(-L(g(u))) \phi(g(u)) |g'(u)| du$$

$$L(g(u)) = n \prod u_i^{2k_i} + n^{1/2} \prod u_i^{k_i} \psi(u)$$

Application to Bayes Estimation



Log marginal likelihood for n samples

$$F(\mathbf{n}) = -\log \int dt \int du \delta(t - \mathbf{n} \Pi u_i^{2k_i}) u_i^{h_i} \exp(-t - t^{1/2} \psi(u)) \phi(g(u))$$

$$\longrightarrow \lambda \log n - (m-1) \log \log n + \text{random variable}$$

Generalization Error

$$\longrightarrow E[G] = \lambda / n - (m-1) / n \log n$$

S. Watanabe, Neural Computation, Vol. 13, No. 4, pp. 899-933, 2001

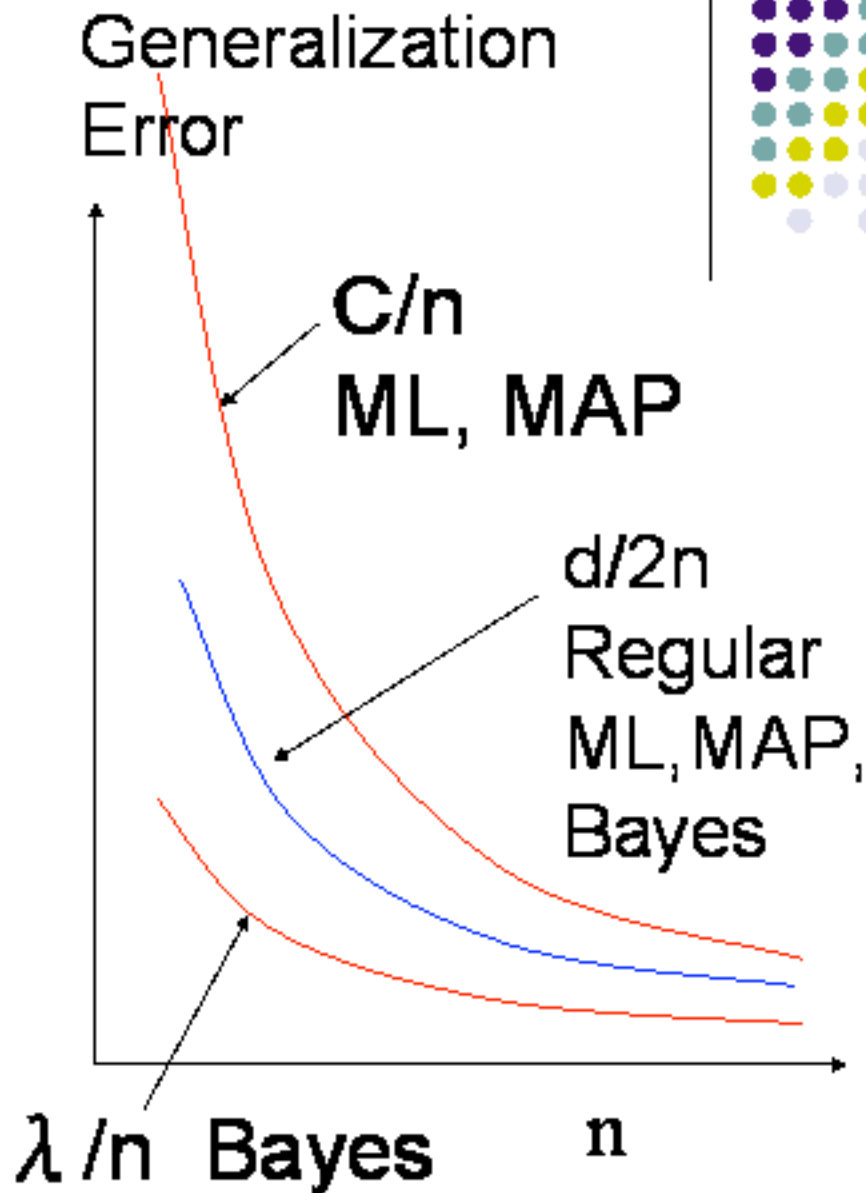


Theorem 2.

The Bayes generalization error is given by

$$E[G] = \lambda/n - (m-1)/n \log n$$

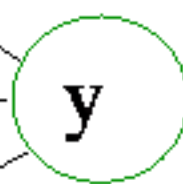
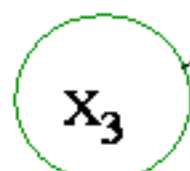
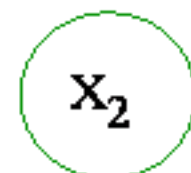
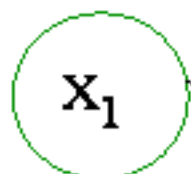
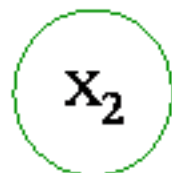
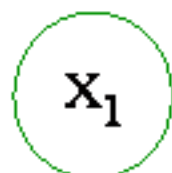
Where $-\lambda$ and m are respectively equal to the largest pole and its order of the zeta function.



A Simple Example



True



Model

$$p(x_1, x_2, x_3 | a, b, c)$$

$$p(x_1, x_2, x_3 | a, b, c) = \frac{1}{Z} \sum_{y=\pm 1} \exp(ax_1 y + bx_2 y + cx_3 y)$$

A Simple Example



$$K(a,b,c) = a^2b^2 + b^2c^2 + c^2a^2$$

$$\left. \begin{aligned} a &= a_1 \\ b &= a_1 b_1 \\ c &= a_1 c_1 \end{aligned} \right\}$$

Blowing-up

$$K(a_1, b_1, c_1) = a_1^4 (b_1^2 + b_1^2 c_1^2 + c_1^2)$$

$$\left. \begin{aligned} b_1 &= b_2 \\ c_1 &= b_2 c_2 \end{aligned} \right\}$$

Blowing-up

$$K(a_2, b_2, c_2) = a_2^4 b_2^2 (1 + b_2^4 c_2^2 + c_2^2)$$

Normal Crossing



A Simple Example

$$K(g(u)) = a_2^4 b_2^2 (1 + b_2^4 c_2^2 + c_2^2)$$

$$|g'(u)| = a_2^2 b_2^1$$

$$\xi(z) = \int K(w)^2 \phi(w) dw$$

$$= \sum \int K(g(u))^2 \phi(g(u)) |g'(u)| du$$

$$= \frac{C_1}{4z+3} + \frac{C_2}{2z+2} + \dots \quad \rightarrow \lambda = 3/4$$
$$G = 3/4n$$

6. Application to Singular Statistics



Reduced rank regression

Aoyagi, Watanabe, *Neural Networks*, Vol.J88-D-II,No.10,pp.2112-2124,2005.

Bayesian networks

Rusakov, Geiger, *J. Machine Learning Research*, Vol.6, No.1, pp.1-35,2005

Neural Networks

Aoyagi, Watanabe, *IEICE Trans*, Vol.J88-D-II,No.10,pp.2112-2124,2005.

Normal mixture

Yamazaki, Watanabe *Neural Networks*, Vol.16, No.7, pp.1029-1038,2003.

Hidden Markov Models

Yamazaki, Watanabe, *Neurocomputing*, Vol.69,pp.62-84,2005

Summary



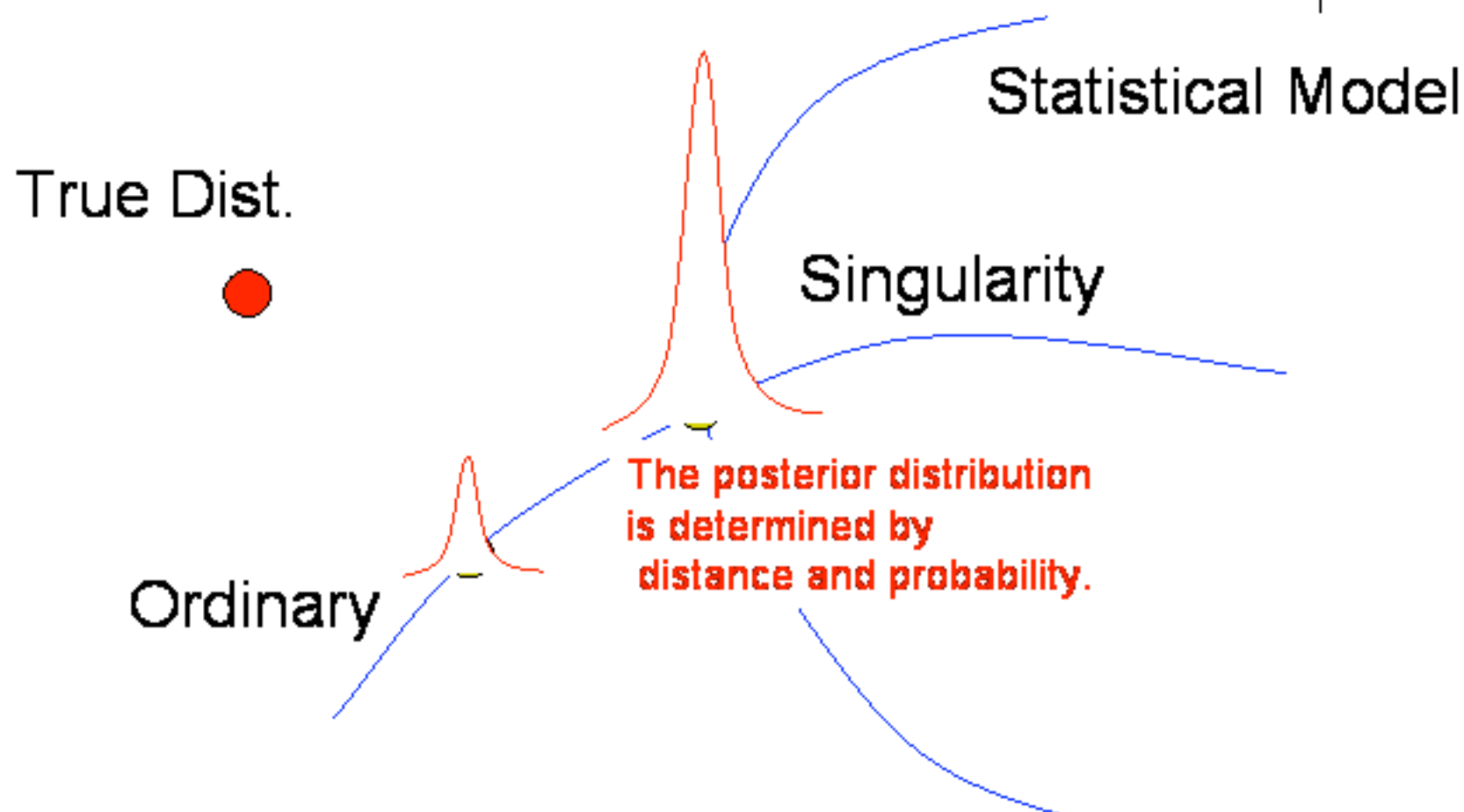
Statistical Models for computational biology is singular.

**Algebraic geometry is needed.
Resolution of Singularities plays
the central role.**

**Applications to ML, MAP, and Bayes
are introduced.**



Q: The case when the true is not on singularity





Q: The case when the true is not on singularity

(1) $\text{Relative_Entropy}(\text{True}, \text{Model}) < C/n$

Theorem holds.

(2) $\text{Relative_Entropy}(\text{True}, \text{Model}) > C/n$

Use the more complex model

(3) $\text{Relative_Entropy}(\text{True}, \text{Model}) = C/n$

Model Selection is needed using the theorem (1).