# What is Singular Learning Theory?

Shaowei Lin (UC Berkeley)

16 April 2012

University of Washington

# Statistical Motivation: 132 Schizophrenic Patients

## Model Selection

Evans-Gilula-Guttman(1989) studied schizophrenic patients for connections between recovery time (in years $Y$) and frequency of visits by relatives.

| | $2 \le Y < 10$ | $10 \le Y < 20$ | $20 \le Y$ | *Totals* |
|---|---|---|---|---|
| Regularly | 43 | 16 | 3 | *62* |
| Rarely | 6 | 11 | 10 | *27* |
| Never | 9 | 18 | 16 | *43* |
| *Totals* | *58* | *45* | *29* | **132** |

They wanted to find out if the data can be explained by the *independence model* or a *naïve Bayes model* with two hidden states (e.g. male and female).

# Model Selection

**Independence Model $\mathcal{M}_{\mathrm{I}}$:**
parametrized by $(a, b) \in \Delta_2 \times \Delta_2$.

| | $2 \leq Y < 10$ | $10 \leq Y < 20$ | $20 \leq Y$ |
|---|---|---|---|
| Regularly | $a_1 b_1$ | $a_1 b_2$ | $a_1 b_3$ |
| Rarely | $a_2 b_1$ | $a_2 b_2$ | $a_2 b_3$ |
| Never | $a_3 b_1$ | $a_3 b_2$ | $a_3 b_3$ |

**Naïve Bayes Model $\mathcal{M}_{\mathrm{NB}}$:**
parametrized by $(t, a, b, c, d) \in \Delta_1 \times \Delta_2 \times \Delta_2 \times \Delta_2 \times \Delta_2$.

| | $2 \leq Y < 10$ | $10 \leq Y < 20$ | $20 \leq Y$ |
|---|---|---|---|
| Regularly | $ta_1 b_1 + (1-t)c_1 d_1$ | $ta_1 b_2 + (1-t)c_1 d_2$ | $ta_1 b_3 + (1-t)c_1 d_3$ |
| Rarely | $ta_2 b_1 + (1-t)c_2 d_1$ | $ta_2 b_2 + (1-t)c_2 d_2$ | $ta_2 b_3 + (1-t)c_2 d_3$ |
| Never | $ta_3 b_1 + (1-t)c_3 d_1$ | $ta_3 b_2 + (1-t)c_3 d_2$ | $ta_3 b_3 + (1-t)c_3 d_3$ |

Because $\mathcal{M}_{\mathrm{I}}$ is a submodel of $\mathcal{M}_{\mathrm{NB}}$, model selection using *maximum likelihood* will always choose $\mathcal{M}_{\mathrm{NB}}$.

We do model selection using the *marginal likelihood* instead.

Schizophrenic Patients
- Model Selection
- Marginal Likelihood
- Exact Evaluation
- Asymptotic Approx
- Sumio Watanabe

Integral Asymtotics

Singular Learning

Algebraic Geometry

Applicatons

## Marginal Likelihood

$$Z_N = \int_\Omega \prod_{i,j} p_{ij}(\omega)^{U_{ij}} \, \varphi(\omega) d\omega$$

| | |
|---|---|
| $U_{ij}, N$ | sample state frequencies, sample size |
| $\omega, \Omega$ | model parameters, parameter space |
| $p_{ij}(\omega)$ | model state probabilities |
| $\varphi(\omega)$ | prior on parameter space |

Generally, evaluating such integrals accurately is a difficult problem. Existing methods can be divided into three broad classes:

1. *Exact* evaluation by closed form formulas
2. *Numerical* estimation by Monte Carlo techniques
3. *Asymptotic* approximation by analyzing large samples

# Exact Evaluation

$$Z_N = \int_\Omega \prod_{i,j} p_{ij}(\omega)^{U_{ij}} \, \varphi(\omega) d\omega$$

In special cases, we can find *closed form formulas* for the integral.

Lin-Sturmfels-Xu(2009) computed this integral for $\mathcal{M}_{\text{NB}}$ *exactly* (not a floating point approx) assuming the uniform prior $\varphi(\omega) = 1$.

It is the rational number with numerator

$$\begin{array}{l} 278019488531063389120643600324989329103876140805 \\ 285242839582092569357265886675322845874097528033 \\ 99493069713103633199906939405711180837568853737 \end{array}$$

and denominator

$$\begin{array}{l} 12288402873591935400678094796599848745442833177572204 \\ 50448819979286456995185542195946815073112429169997801 \\ 33503900169921912167352239204153786645029153951176422 \\ 43298328046163472261962028461650432024356339706541132 \\ 34375318471880274818667657423749120000000000000000. \end{array}$$

## Asymptotic Approximation

$$Z_N = \int_\Omega \prod_{i,j} p_{ij}(\omega)^{U_{ij}} \varphi(\omega) d\omega$$

Study the behavior of the integral as sample size $N$ grows large.

$$U = Nq, \quad q = \frac{1}{132}\begin{pmatrix} 43 & 16 & 3 \\ 6 & 11 & 10 \\ 9 & 18 & 16 \end{pmatrix}, \quad q = \frac{1}{132}\begin{pmatrix} 43.00 & 16.00 & 3.00 \\ 5.98 & 11.12 & 9.90 \\ 9.02 & 17.88 & 16.10 \end{pmatrix}$$

Different asymptotic directions (*true distributions*) $q$ for the data may give different asymptotic approximations.

- Bayesian Information Criterion

$$-\log Z_N \approx \text{BIC} = -\sum_{i,j} U_{ij} \log q_{ij} + \frac{d}{2} \log N$$

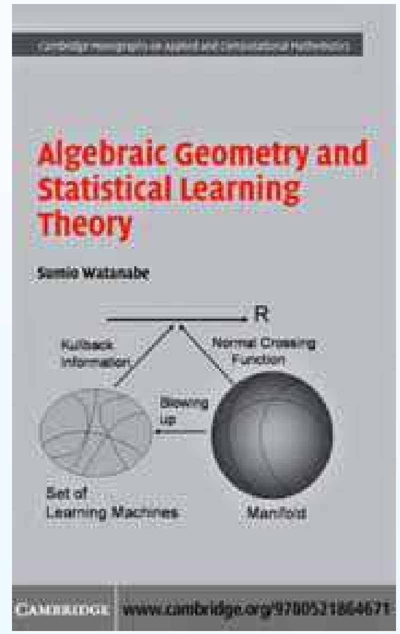
where $d$ is the dimension of the parameter space.

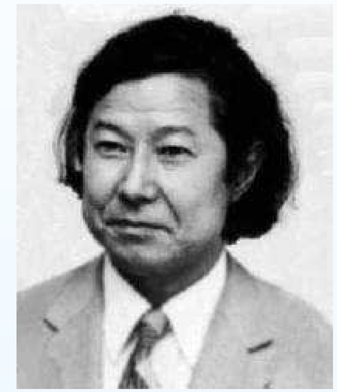
The BIC holds for *smooth* models (e.g. multinomial, exponential) but generalization to *singular* models (e.g. hidden variables) unknown.

# Sumio Watanabe



Sumio Watanabe



Heisuke Hironaka

In 1998, Sumio Watanabe discovered how to study the asymptotic behavior of singular models. His insight was to use a deep result in algebraic geometry known as Hironaka's Resolution of Singularities.

Heisuke Hironaka proved this celebrated result in 1964.
His accomplishment won him the Field's Medal in 1970.

## Asymptotic Approximation

$$Z_N = \int_\Omega \prod_{i,j} p_{ij}(\omega)^{U_{ij}} \varphi(\omega) d\omega$$

Using Watanabe's *Singular Learning Theory*,

$$-\log Z_N \approx -\sum_{i,j} U_{ij} \log q_{ij} + \lambda_q \log N - (\theta_q - 1) \log\log N$$

where the *learning coefficient* $(\lambda_q, \theta_q)$ is given by

$$(\lambda_q, \theta_q) = \begin{cases} (5/2, 1) & \text{if } \operatorname{rank} q = 1, \\ (7/2, 1) & \text{if } \operatorname{rank} q = 2,\ q \notin \left[\begin{smallmatrix} 0 & \times \\ \times & \times \end{smallmatrix}\right] \cup \left[\begin{smallmatrix} 0 & \times \\ \times & 0 \end{smallmatrix}\right], \\ (4, 1) & \text{if } \operatorname{rank} q = 2,\ q \in \left[\begin{smallmatrix} 0 & \times \\ \times & \times \end{smallmatrix}\right] \setminus \left[\begin{smallmatrix} 0 & \times \\ \times & 0 \end{smallmatrix}\right], \\ (9/2, 1) & \text{if } \operatorname{rank} q = 2,\ q \in \left[\begin{smallmatrix} 0 & \times \\ \times & 0 \end{smallmatrix}\right]. \end{cases}$$

Here, $q \in \left[\begin{smallmatrix} 0 & \times \\ \times & \times \end{smallmatrix}\right]$ if for some $i, j$, $q_{ii} = 0$ and $q_{ij}\, q_{ji}\, q_{jj} \neq 0$,
$q \in \left[\begin{smallmatrix} 0 & \times \\ \times & 0 \end{smallmatrix}\right]$ if for some $i, j$, $q_{ii} = q_{jj} = 0$ and $q_{ij}\, q_{ji} \neq 0$.

# Mathematical Technique:
# Integral Asymptotics

# Integral Asymptotics

For large $N$, approximate

$$Z(N) \;=\; \int_{[0,1]^2} (1 - x^2y^2)^{N/2}\, dxdy.$$

- Write $Z(N)$ as $\int e^{-Nf(x,y)} dxdy$ where

$$f(x,y) = -\frac{1}{2}\log(1 - x^2y^2).$$

- Can we use the Gaussian integral

$$\int_{\mathbb{R}^d} e^{-\frac{N}{2}(\omega_1^2+\cdots+\omega_d^2)} d\omega = \left(\frac{2\pi}{N}\right)^{d/2}$$

  by finding a suitable change of coordinates for $x, y$?

## Laplace Approximation

$\Omega$ small nbhd of origin, $f : \Omega \to \mathbb{R}$ analytic function with unique minimum $f(0)$ at origin, $\partial^2 f$ Hessian of $f$. If $\det \partial^2 f(0) \neq 0$,

$$Z(N) \;=\; \int_\Omega e^{-Nf(\omega)} d\omega \;\approx\; e^{-Nf(0)} \cdot \sqrt{\frac{(2\pi)^d}{\det \partial^2 f(0)}} \cdot N^{-d/2}.$$

- e.g. Bayesian Information Criterion

$$-\log Z_N \;\approx\; \text{BIC} \;=\; \Big(-\sum_{i,j} U_{ij} \log q^*_{ij}\Big) + \frac{d}{2} \log N$$

- e.g. Stirling's approximation

$$N! \;=\; N^{N+1} \int_0^\infty e^{-N(x - \log x)}\, dx \;\approx\; N^{N+1} e^{-N} \sqrt{\frac{2\pi}{N}}$$

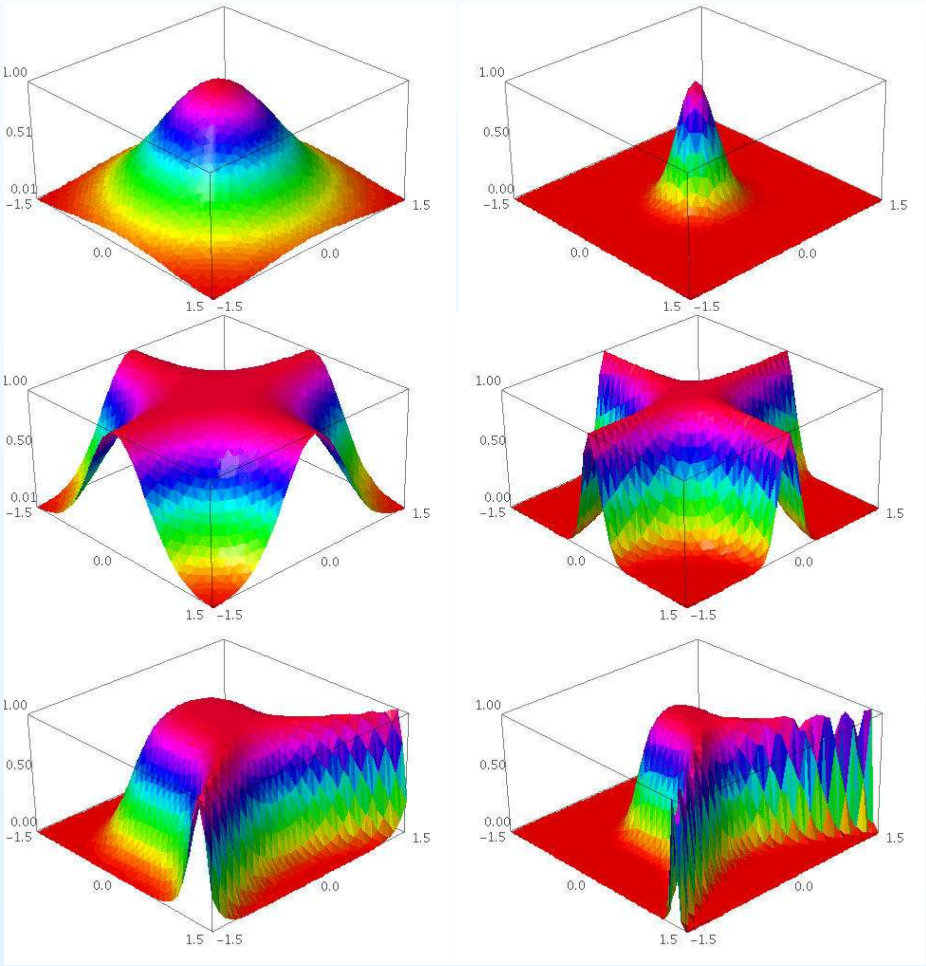
## Geometry of the Integral

Because $\det \partial^2 f(0) = 0$ in our example, we cannot apply Laplace approximation. More important to study *minimas* of $f$.

$$f(x, y) = x^2 + y^2$$

$$f(x, y) = (xy)^2$$

$$f(x, y) = (y^2 - x^3)^2$$



Plots of $z = e^{-Nf(x,y)}$ for $N = 1$ and $N = 10$

## Monomial Functions



Notation: $\omega^\kappa = \omega_1^{\kappa_1} \cdots \omega_d^{\kappa_d}$.

Asymptotic theory of Arnol'd, Guseĭn-Zade and Varchenko (1974).

**Theorem (AGV).** Given $\kappa, \tau \in \mathbb{Z}^d_{\geq 0}$,

$$Z(N) \;=\; \int_\Omega e^{-N\omega^\kappa} \omega^\tau d\omega \;\approx\; CN^{-\lambda} (\log N)^{\theta-1}$$

where $\Omega \subset \mathbb{R}^d$ is a compact nbhd of the origin, $C$ is a constant,

$$\lambda = \min_i \frac{\tau_i + 1}{\kappa_i},$$

$\theta =$ number of times minimum is attained.

# Resolution of Singularities

Let $\Omega \subset \mathbb{R}^d$ and $f : \Omega \to \mathbb{R}$ analytic function.

- We say $\rho : U \to \Omega$ *desingularizes* $f$ if

  1. $U$ is a $d$-dimensional real analytic manifold covered by patches $U_1, \ldots, U_s$ ($\simeq$ subsets of $\mathbb{R}^d$).

  2. For each restriction $\rho : U_i \to \Omega$, $\mu \mapsto \omega$,

  $$f \circ \rho(\mu) = a(\mu)\mu^\kappa, \quad \det \rho'(\mu) = b(\mu)\mu^\tau$$

  where $a(\mu)$ and $b(\mu)$ are nonzero on $U_i$.

- Hironaka (1964) proved that desingularizations always exist.
- The preimage (*transform*) $\{\mu : f \circ \rho(\mu) = 0\}$ of the zero-set (*variety*) $\{\omega : f(\omega) = 0\}$ has *simple normal crossings*.

# Algorithm for Computing Integral Asymptotics

$$Z(N) \;=\; \int_\Omega e^{-Nf(\omega)}\varphi(\omega)d\omega \;\approx\; e^{-Nf^*}\cdot CN^{-\lambda}(\log N)^{\theta-1}$$

**Input**:

Semialgebraic set $\Omega = \{\omega : g_1(\omega) \geq 0, \ldots, g_l(\omega) \geq 0\} \subset \mathbb{R}^d$

Analytic functions $f, \varphi : \Omega \to \mathbb{R}$

**Output**:

Asymptotic coefficients $f^*, \lambda, \theta$

1. Find minimum $f^*$ of $f$ over $\Omega$.
2. Find a desingularization $\rho$ for product $(f - f^*)g_1 \cdots g_l\varphi$.
3. Use AGV Theorem to find coefficients $\lambda_i, \theta_i$ on each patch $U_i$.
4. $\lambda = \min\{\lambda_i\}$, $\theta = \max\{\theta_i : \lambda_i = \lambda\}$.

# Higher Order Asymptotics

After desingularizing $f(x,y) = -\frac{1}{2}\log(1-x^2y^2)$, we were able to compute higher order asymptotics of $Z(N)$.

$$\begin{aligned}
&\sqrt{\frac{\pi}{8}}N^{-\frac{1}{2}}\log N && -\sqrt{\frac{\pi}{8}}\left(\frac{1}{\log 2} - 2\log 2 - \gamma\right)N^{-\frac{1}{2}} \\
&-\frac{1}{4}N^{-1}\log N && +\frac{1}{4}\left(\frac{1}{\log 2} + 1 - \gamma\right)N^{-1} \\
&-\frac{\sqrt{2\pi}}{128}N^{-\frac{3}{2}}\log N && +\frac{\sqrt{2\pi}}{128}\left(\frac{1}{\log 2} - 2\log 2 - \frac{10}{3} - \gamma\right)N^{-\frac{3}{2}} \\
& && -\frac{1}{24}N^{-2} \quad + \quad \cdots
\end{aligned}$$

Euler-Mascheroni constant

$$\gamma = \lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k} - \log n\right) \approx 0.5772156649.$$

# Singular Learning Theory

# Statistical Model

| | |
|---|---|
| $X$ | random variable with state space $\mathcal{X}$ (e.g. $\{1, 2, \ldots, k\}$, $\mathbb{R}^k$) |
| $\Delta_{\mathcal{X}}$ | space of probability distributions on $\mathcal{X}$ |

| | |
|---|---|
| $\mathcal{M} \subset \Delta_{\mathcal{X}}$ | statistical model, image of $p : \Omega \to \Delta_{\mathcal{X}}$ |
| $\Omega$ | parameter space |
| $p(x\|\omega)dx$ | distribution at $\omega \in \Omega$ |
| $\varphi(\omega)d\omega$ | prior distribution on $\Omega$ |

Given samples $X_1, \ldots, X_N$ of $X$, define *marginal likelihood*

$$Z_N = \int_\Omega \prod_{i=1}^{N} p(X_i|\omega)\, \varphi(\omega) d\omega.$$

Given $q \in \Delta_{\mathcal{X}}$, define *Kullback-Leibler function*

$$K(\omega) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x|\omega)} dx.$$

# Learning Coefficient

Suppose samples $X_1, \ldots, X_N$ are drawn from distribution $q \in \mathcal{M}$.
Define *empirical entropy* $S_N = -\frac{1}{N}\sum_{i=1}^N \log q(X_i)$.

**Convergence of stochastic complexity (Watanabe)**

The *stochastic complexity* has the asymptotic expansion

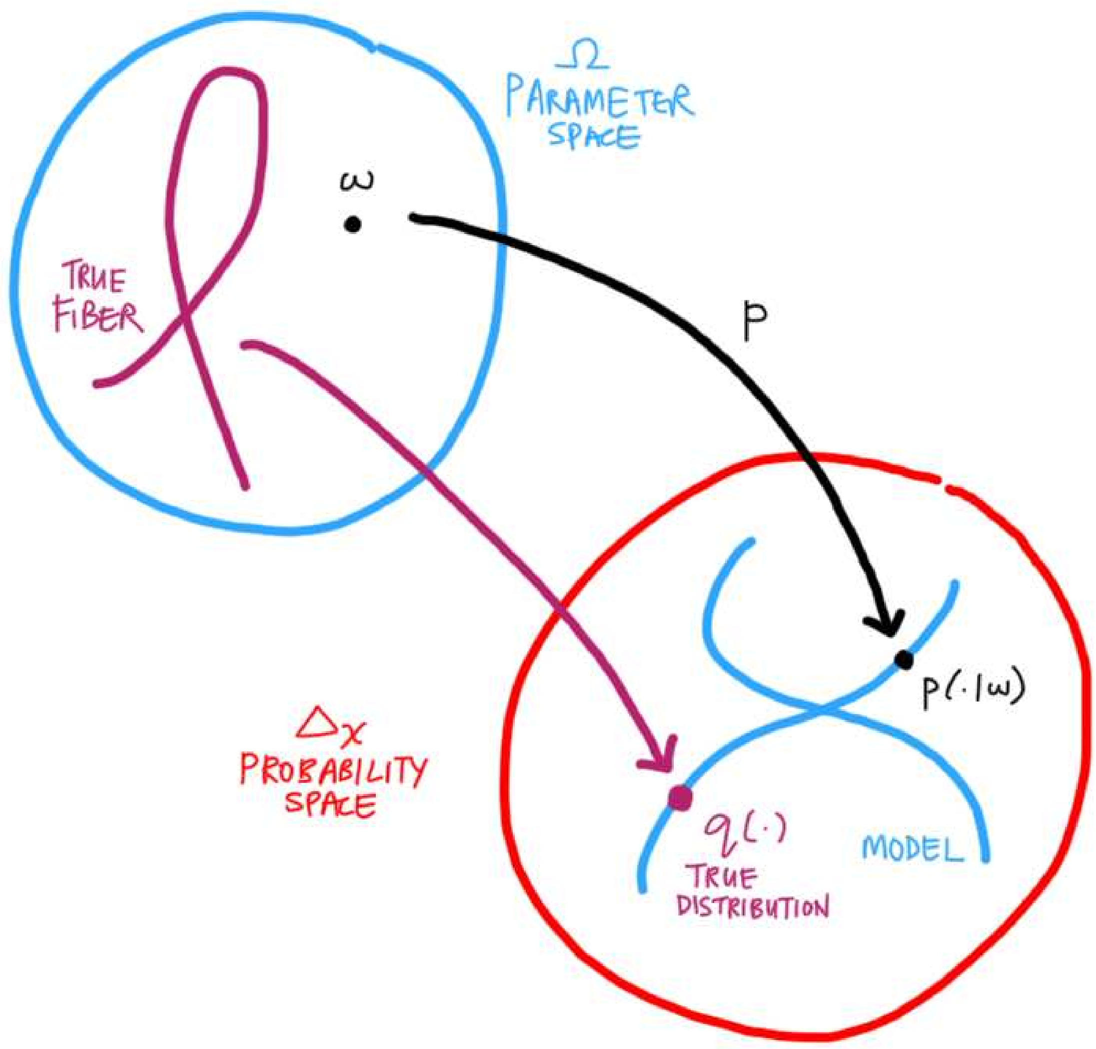
$$-\log Z_N = NS_N + \lambda_q \log N - (\theta_q - 1)\log\log N + R_N$$

where $R_N$ converges in law to a random variable. Moreover, $\lambda_q, \theta_q$ are asymptotic coefficients of the deterministic integral

$$Z(N) = \int_\Omega e^{-NK(\omega)}\varphi(\omega)d\omega \;\approx\; CN^{-\lambda_q}(\log N)^{\theta_q - 1}.$$

Think of this as a *Bayesian Information Criterion* for singular models.
$(\lambda_q, \theta_q)$ is the *learning coefficient* of the model $\mathcal{M}$ at $q$.

# Geometry of Singular Models



# Standard Form of Log Likelihood Ratio

Define *log likelihood ratio*. Note that its expectation is $K(\omega)$.

$$K_N(\omega) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{q(X_i)}{p(X_i|\omega)}.$$

**Standard Form of Log Likelihood Ratio (Watanabe)**

If $\rho : U \to \Omega$ desingularizes $K(\omega)$, then on each patch $U_i$,

$$K_N \circ \rho(\mu) = \mu^{2\kappa} - \frac{1}{\sqrt{N}}\mu^{\kappa}\xi_N(\mu)$$

where $\xi_N(\mu)$ converges in law to a Gaussian process on $U$.

Think of this as a *Central Limit Theorem* for singular models.

## Bayes Generalization Error



Given samples $D = \{X_1, \ldots, X_N\}$, define

$$p(\omega|D) = \text{posterior distribution} = \tfrac{1}{Z_N}\varphi(\omega)\textstyle\prod_{i=1}^N p(X_i|\omega)$$

$$p(x|D) = \text{predictive distribution} = \textstyle\int_\Omega p(x|\omega)p(\omega|D)d\omega$$

The *Bayes Generalization Error* $B_N$ is the Kullback-Leibler distance from the true distribution $q(x)$ to the predictive distribution $p(x|D)$.

$$B_N = \int_{\mathcal{X}} q(x)\log\frac{q(x)}{p(x|D)}dx$$

Let $\hat{\omega}$ denote the MLE. *Asymptotically*, $B_N$ is equivalent to

- Akaike Information Criterion

$$\text{AIC} = -\textstyle\sum_{i=1}^N \log p(X_i|\hat{\omega}) + d$$

- Akaike Information Criterion for singular models

$$\text{AIC} = -\textstyle\sum_{i=1}^N \log p(X_i|\hat{\omega}) + 2\nu_q$$

where $\nu_q$ is the *singular fluctuation*.

## Bayes Generalization Error

Let $\mathbb{E}_X$ denote expectation over the data distribution.
Let $\mathbb{E}_w$ denote expectation over the posterior distribution $p(\omega|D)$.
Given a function $f(\omega)$, we can numerically estimate:
$\mathbb{E}_w[f(\omega)]$ by sampling from $p(\omega|D)$ using MCMC methods,
$\mathbb{E}_X[f(X)]$ by averaging $f(X_i)$ over the data $X_1, \ldots, X_N$.

*Numerical* estimates of $B_N$:

- Deviance Information Criterion

$$\text{DIC} = \mathbb{E}_X[\log p(X|\mathbb{E}_\omega[\omega])] - 2\,\mathbb{E}_\omega[\mathbb{E}_X[\log p(X|\omega)]]$$

- Widely Applicable Information Criterion for singular models

$$\text{WAIC} = \mathbb{E}_X[\log \mathbb{E}_\omega[p(X|\omega)]] - 2\,\mathbb{E}_\omega[\mathbb{E}_X[\log p(X|\omega)]]$$

# Mathematical Questions in Singular Learning



For each distribution $q$ in the model $\mathcal{M}$,

1. Study the geometrical structure of the fiber $p^{-1}(q)$.
2. Study the asymptotics of the integral

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

   and compute the learning coefficient $(\lambda_q, \theta_q)$.
3. Desingularize the Kullback-Leibler function $K(\omega)$.

# Computations: Algebraic Geometry

## Ideals & Varieties

Polynomial system (*ideal*) $\langle y - x^2, y \rangle \subset \mathbb{R}[x, y]$ $\longleftrightarrow$ Solution set (*variety*) $V = \{(0,0)\} \subset \mathbb{R}^2$

- If $y - x^2$ and $y$ vanish on $V$, so do all polynomials of the form
$$p(x,y) = (y - x^2)\, p_1(x,y) + (y)\, p_2(x,y).$$
This infinite set of polynomials is the *ideal* $I = \langle y - x^2, y \rangle$.

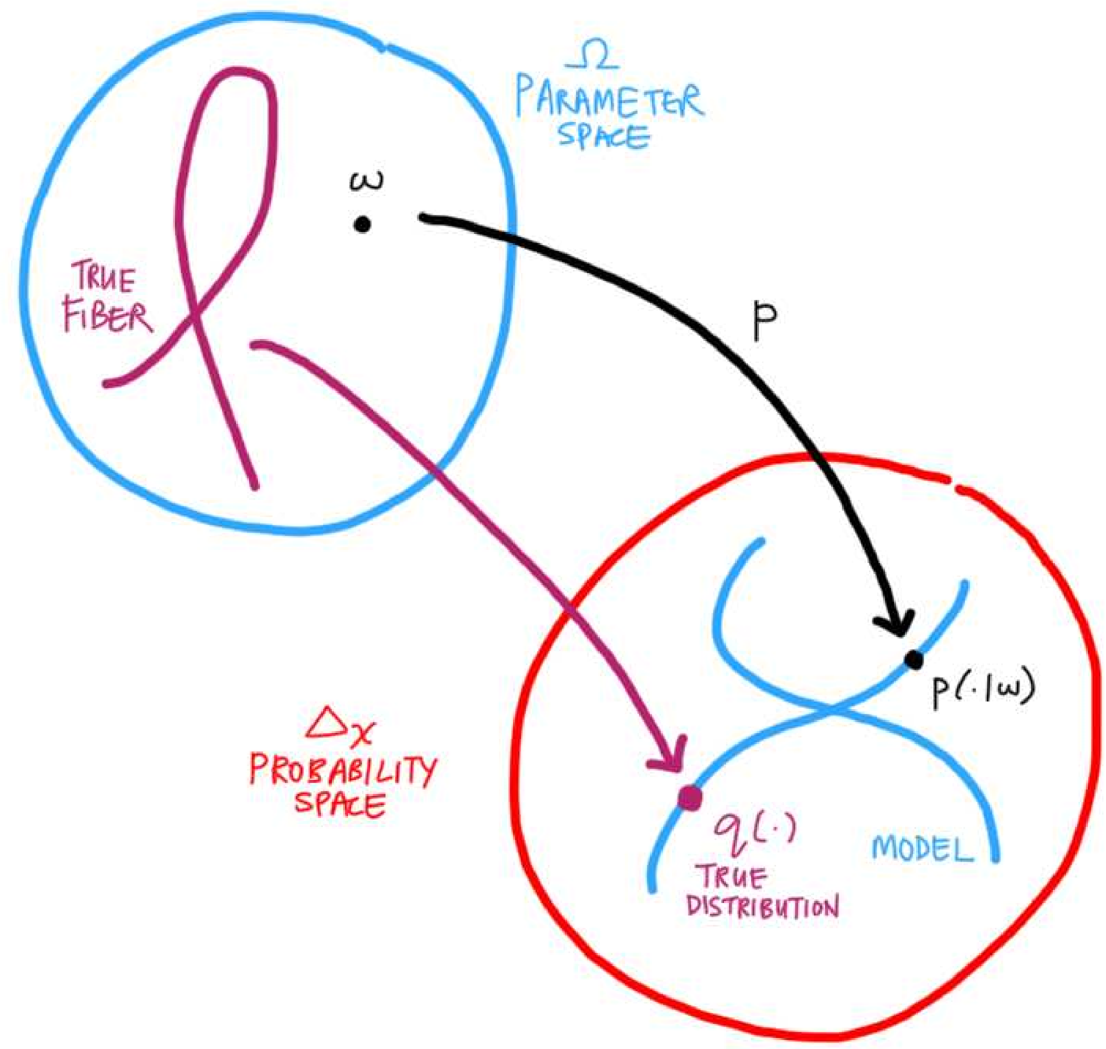
- Vector spaces: generated by addition, scalar multiplication.
Ideals: generated by addition, polynomial multiplication.
Different sets of polynomials can generate the same ideal.

- Given subset $I \subset \mathcal{R} := \mathbb{R}[x_1, \ldots, x_d]$, define the *variety*
$$\mathcal{V}(I) = \{x \in \mathbb{R}^d : f(x) = 0 \text{ for all } f \in I\}.$$
Given subset $V \subset \mathbb{R}^d$, define the *ideal*
$$\mathcal{I}(V) = \{f \in \mathcal{R} : f(x) = 0 \text{ for all } x \in V\}.$$

## Gröbner Bases

- Every system of linear equations has a *row echelon form*, which depends on the ordering of the coordinates and is computed using *Gaussian elimination*.
- Every system of polynomial equations has a *Gröbner basis*, which depends on the ordering of the monomials and is computed using *Buchberger's algorithm*.
- Determine ideal membership, dimension, degree, solutions, irreducible components, elimination of variables, etc. Also essential in resolution of singularities.
- **Textbook**:
  "Ideals, Varieties, and Algorithms," Cox-Little-O'Shea (1997)
  **Software**:
  `Macaulay2`, `Singular`, `Maple`, etc.

# Geometry of Singular Models



# Fiber Ideals

- **Discrete Models.** State probabilities parametrized by

$$p(\omega) \in \Delta_{k-1}.$$

Given a true distribution $\hat{p}$ that lies in the model, define the fiber ideal of the model at $\hat{p}$ to be

$$I_{\hat{p}} = \langle p_1(\omega) - \hat{p}_1, \ldots, p_k(\omega) - \hat{p}_k \rangle.$$

- **Gaussian Models.** Mean and covariance parametrized by

$$\mu(\omega) \in \mathbb{R}^k, \quad \Sigma(\omega) \in \mathbb{R}^{k\times k}_{\succ 0}.$$

Given a true distribution $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$ that lies in the model, define the fiber ideal of the model at $(\hat{\mu}, \hat{\Sigma})$ to be

$$\begin{aligned} I_{\hat{\mu},\hat{\Sigma}} = \langle\, & \mu_1(\omega) - \hat{\mu}_1, \ldots, \mu_k(\omega) - \hat{\mu}_k, \\ & \Sigma_{11}(\omega) - \hat{\Sigma}_{11}, \ldots, \Sigma_{kk}(\omega) - \hat{\Sigma}_{kk} \,\rangle. \end{aligned}$$

## Real Log Canonical Thresholds

Given ideal $I = \langle f_1(\omega), \ldots, f_k(\omega) \rangle \subset \mathbb{R}[\omega_1, \ldots, \omega_d]$,
polynomial $\varphi(\omega) \in \mathbb{R}[\omega_1, \ldots, \omega_d]$,
semialgebraic set $\Omega \subset \mathbb{R}^d$ with boundary eqns $g_1, \ldots, g_l$.

The *real log canonical threshold* $(\lambda, \theta)$ of $I$ at $x \in \Omega$ satisfies

$$\int_{\Omega_x} e^{-N(f_1^2+\cdots+f_k^2)} \varphi(\omega) d\omega \;\approx\; C N^{-\lambda} (\log N)^{\theta - 1}$$

for suff small nbhd $\Omega_x$ of $x$ in $\Omega$. Denote $(\lambda, \theta) = \mathrm{RLCT}_{\Omega_x}(I; \varphi)$.

### Properties

- Definition is independent of choice of generators $f_1, \ldots, f_k$.
- $\lambda$ positive *rational* number, $\theta$ positive *integer*.
- Order the $(\lambda, \theta)$ by the value of $N^{\lambda} (\log N)^{-\theta}$ for large $N$.
- Depends on structure of boundary $\partial\Omega$ if $x \in \partial\Omega$.

# Real Log Canonical Thresholds

Suppose we have a discrete or Gaussian model with parameter space $\Omega$ and prior $\varphi(\omega)$, and a true distribution $q$ in the model.

**Theorem (L.)**

The learning coefficient $(\lambda_q, \theta_q)$ is given by

$$(2\lambda_q, \theta_q) = \min_{x \in \mathcal{V}(I_q)} \mathrm{RLCT}_{\Omega_x}(I_q; \varphi)$$

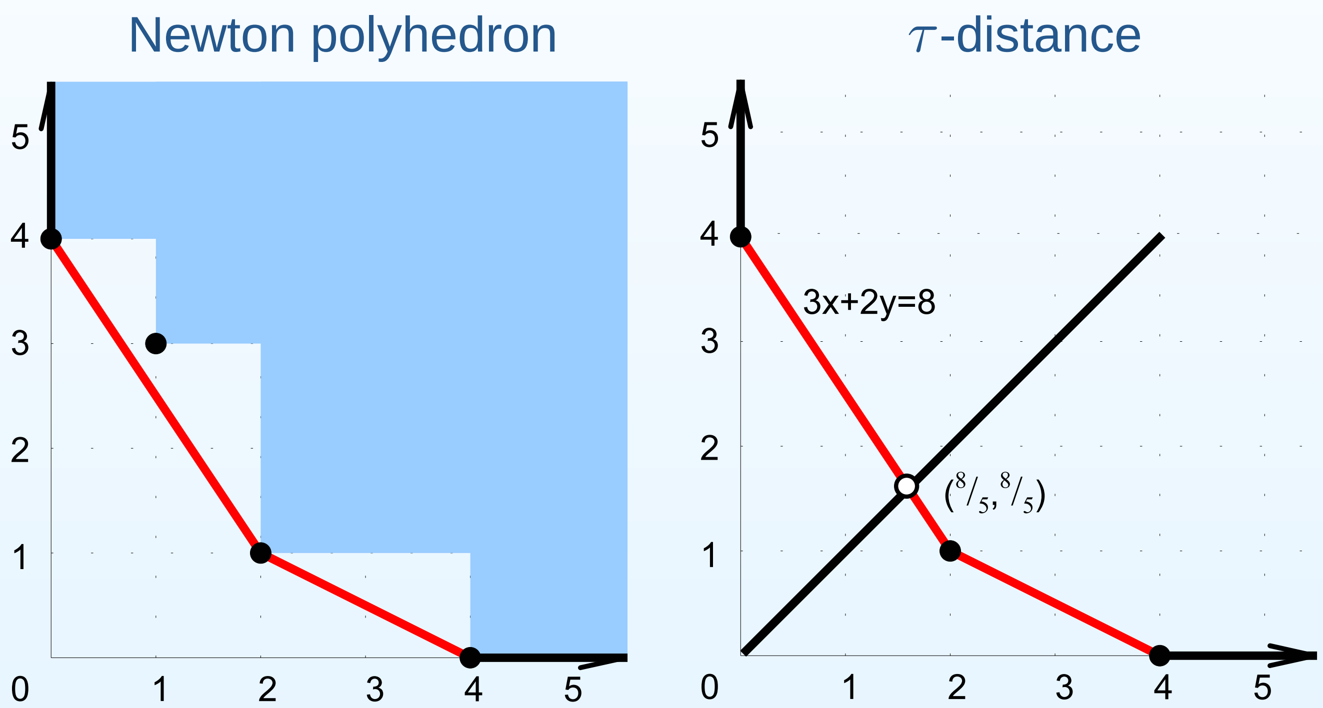
where $I_q$ is the fiber ideal at $q$ and $\mathcal{V}(I_q) \subset \Omega$ is the fiber over $q$.

**Algorithm for Computing $(\lambda, \theta) = \mathrm{RLCT}_{\Omega_x}(I; \varphi)$**

1. Shift the origin to $x$.
2. Find *monomialization* $\rho : U \to \Omega$ for $I, g_1, \ldots, g_l, \varphi$.
   (Transform of $I$ is generated by *monomials* on each patch $U_i$)
3. Find RLCT $(\lambda_i, \theta_i)$ on each patch $U_i$ using *Newton polyhedra*.
4. $\lambda = \min\{\lambda_i\}$, $\theta = \max\{\theta_i : \lambda_i = \lambda\}$.

# Newton Polyhedra

e.g. Let $I = \langle x^4, x^2y, xy^3, y^4 \rangle$ and $\tau = (0,0)$.

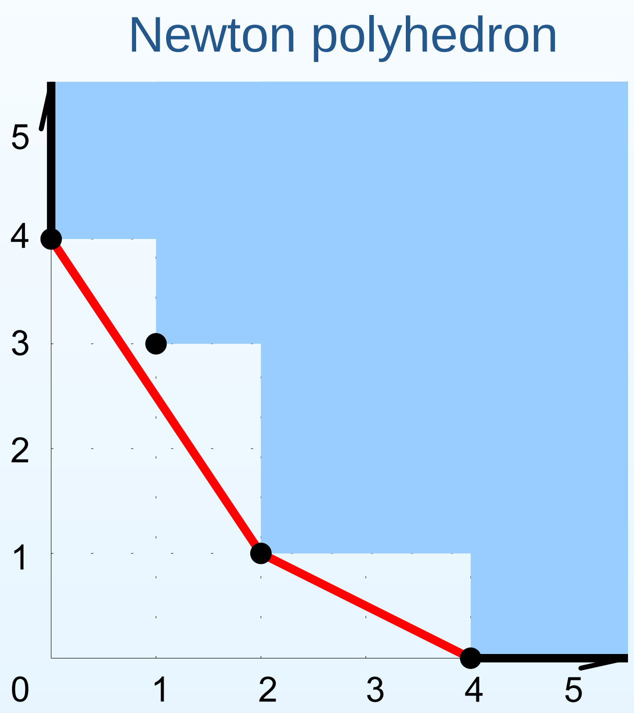


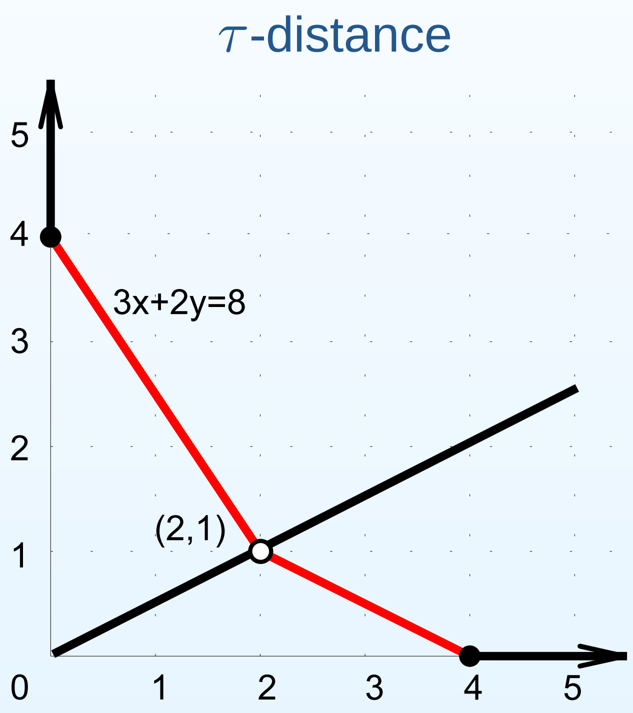

The $\tau$-distance is $l_\tau = 8/5$ and the multiplicity is $\theta_\tau = 1$.

# Newton Polyhedra



e.g. Let $I = \langle x^4, x^2y, xy^3, y^4 \rangle$ and $\tau = (1, 0)$.




The $\tau$-distance is $l_\tau = 1$ and the multiplicity is $\theta_\tau = 2$.

# Newton Polyhedra

Given an ideal $I \subset \mathbb{R}[\omega_1, \ldots, \omega_d]$,

1. Plot $\alpha \in \mathbb{R}^d$ for each monomial $\omega^\alpha$ appearing in some $f \in I$.
2. Take the convex hull $\mathcal{P}(I)$ of all plotted points.

This convex hull $\mathcal{P}(I)$ is the *Newton polyhedron* of $I$.

Given a vector $\tau \in \mathbb{Z}^d_{\geq 0}$, define

1. *$\tau$-distance* $l_\tau = \min\{t : t(\tau_1 + 1, \ldots, \tau_d + 1) \in \mathcal{P}(I)\}$.
2. *multiplicity* $\theta_\tau =$ codim of face of $\mathcal{P}(I)$ at this intersection.

## Upper Bounds

Let $\Omega \subset \mathbb{R}^d$ be a sufficiently small nbhd of the origin.

**Proposition (Trivial)** $\quad \mathrm{RLCT}_\Omega(I;\varphi) \leq d$

**Theorem (Watanabe)** $\quad \mathrm{RLCT}_\Omega(I;\varphi) \leq \operatorname{codim} \mathcal{V}(I)$

**Theorem (L.)**

If $l_\tau$ is the $\tau$-distance of $\mathcal{P}(I)$ and $\theta_\tau$ is its multiplicity, then

$$\mathrm{RLCT}_\Omega(I;\omega^\tau) \leq (1/l_\tau, \theta_\tau).$$

Equality occurs when $I$ is a monomial ideal.

# Applications to Statistics

# Bayesian Information Criterion

When the model is regular, the fiber ideal is $I = \langle \omega_1, \ldots, \omega_d \rangle$.
Using Newton polyhedra, the RLCT of this ideal is $(d, 1)$.

By our theorem, the learning coefficient is $(\lambda, \theta) = (d/2, 1)$.
By Watanabe's theorem, asymptotically

$$-\log Z_N \;\approx\; -\sum_{i=1}^{N} \log q(X_i) + \frac{d}{2} \log N.$$

This formula is the Bayesian Information Criterion (BIC).

## Coin Toss

$$Z(N) \;=\; \int_{[0,1]^2} (1-x^2y^2)^{N/2}\, dxdy.$$

The integral $Z(N)$ comes from the coin toss model parametrized by

$$\begin{aligned} p_1(\omega,t) &= \tfrac{1}{2}t + (1-t)\,\omega &&= \tfrac{1}{2}(1+xy) \\ p_2(\omega,t) &= \tfrac{1}{2}t + (1-t)(1-\omega) &&= \tfrac{1}{2}(1-xy) \end{aligned}$$

where we substituted $\omega = (1+x)/2, t = 1-y$.

Here, the true distribution is $\hat{p}_1 = \hat{p}_2 = 1/2$ and the fiber ideal is

$$I_{\hat{p}} = \langle \tfrac{1}{2}(1+xy) - \tfrac{1}{2}, \tfrac{1}{2}(1-xy) - \tfrac{1}{2} \rangle = \langle xy \rangle.$$

Using Newton polyhehra with $\tau = (0,0)$, we have $(l_\tau, \theta_\tau) = (1,2)$.

Therefore, the RLCT is $(1,2)$, the learning coefficient is $(\frac{1}{2}, 1)$, and

$$Z(N) \;\approx\; CN^{-\frac{1}{2}} (\log N)$$

for some constant $C > 0$.

# 132 Schizophrenic Patients



Model parametrized in $\omega = (t, a_1, a_2, \ldots, d_3)$ by

| | $2 \leq Y < 10$ | $10 \leq Y < 20$ | $20 \leq Y$ |
|---|---|---|---|
| Regularly | $ta_1b_1 + (1-t)c_1d_1$ | $ta_1b_2 + (1-t)c_1d_2$ | $ta_1b_3 + (1-t)c_1d_3$ |
| Rarely | $ta_2b_1 + (1-t)c_2d_1$ | $ta_2b_2 + (1-t)c_2d_2$ | $ta_2b_3 + (1-t)c_2d_3$ |
| Never | $ta_3b_1 + (1-t)c_3d_1$ | $ta_3b_2 + (1-t)c_3d_2$ | $ta_3b_3 + (1-t)c_3d_3$ |

Let the true distribution be $\hat{p}_{ij} = \frac{1}{9}$ for all $i, j$.
Consider the point $\hat{w} = (\frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \ldots, \frac{1}{3})$ on the fiber over $\hat{p}$.
Let us compute the RLCT at $\hat{\omega}$ of the fiber ideal

$$I = \langle p_{11}(\omega) - \hat{p}, \ldots, p_{33}(\omega) - \hat{p} \rangle.$$

Using `Macaulay2` and our library `asymptotics.m2`,
we manipulate the ideal and show that

$$\mathrm{RLCT}_{\hat{\omega}}(I; 1) = (6, 2).$$

All the learning coefficients can be computed in this fashion.

# 132 Schizophrenic Patients



We want to approximate the marginal likelihood $Z_N$ of the data

$$\begin{pmatrix} 43 & 16 & 3 \\ 6 & 11 & 10 \\ 9 & 18 & 16 \end{pmatrix}.$$

The EM algorithm gives us the *maximum likelihood distribution*

$$q = \frac{1}{132}\begin{pmatrix} 43.002 & 15.998 & 3.000 \\ 5.980 & 11.123 & 9.897 \\ 9.019 & 17.879 & 16.102 \end{pmatrix}.$$

Using the ML distribution as the *true distribution*, the learning coefficient is $(\frac{7}{2}, 1)$ (compare with $(\frac{9}{2}, 1)$ for BIC).

| | $-\log Z_N$ |
|---|---|
| Exact | $273.1911759$ |
| BIC | $278.3558034$ |
| RLCT | $275.9144024$ |

# Model Selection (Joint work with Russell Steele)

**Question**: The learning coefficients $(\lambda_q, \theta_q)$ of a statistical model $\mathcal{M}$ depend on the true distribution $q$ of the data which is unknown. How do we use these coefficients for model selection?

**Proposal**: The ML criterion and BIC may be expressed as:

$$\text{ML} = \max_{q \in \mathcal{M}} \{-\textstyle\sum_{i=1}^{N} \log q(X_i)\},$$

$$\text{BIC} = \max_{q \in \mathcal{M}} \{-\textstyle\sum_{i=1}^{N} \log q(X_i) + \frac{d}{2} \log N\}.$$

For singular models, the BIC naturally generalizes to

$$\max_{q \in \mathcal{M}} \{-\textstyle\sum_{i=1}^{N} \log q(X_i) + \lambda_q \log N - (\theta_q - 1) \log \log N\}.$$

**Conjecture**: The generalized BIC for singular models is consistent.

"Algebraic Methods for Evaluating Integrals in Bayesian Statistics"

`http://math.berkeley.edu/~shaowei/swthesis.pdf`

(PhD dissertation, May 2011)

## References

1. V. I. Arnol'd, S. M. Guseĭn-Zade and A. N. Varchenko: *Singularities of Differentiable Maps*, Vol. II, Birkhäuser, Boston, 1985.
2. A. Bravo, S. Encinas and O. Villamayor: A simplified proof of desingularisation and applications, *Rev. Math. Iberoamericana* **21** (2005) 349–458.
3. D. A. Cox, J. B. Little, and D. O'Shea: *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, Springer-Verlag, New York, 1997.
4. M. Evans, Z. Gilula and I. Guttman: Latent class analysis of two-way contingency tables by Bayesian methods, *Biometrika* **76** (1989) 557–563.
5. 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.
6. S. Lin, B. Sturmfels and Z. Xu: Marginal likelihood integrals for mixtures of independence models, *J. Mach. Learn. Res.* **10** (2009) 1611–1631.
7. S. Lin: Algebraic methods for evaluating integrals in Bayesian statistics, PhD dissertation, Dept. Mathematics, UC Berkeley (2011).
8. S. Watanabe: *Algebraic Geometry and Statistical Learning Theory*, Cambridge Monographs on Applied and Computational Mathematics **25**, Cambridge University Press, Cambridge, 2009.