Singular Learning Theory: Insights into Model Choice

Martyn Plummer

AppliBUGS, 26 June 2012
What is singular learning theory?


- The interface between algebraic geometry and statistics
- Allows us to consider statistical models as geometrical objects
- Provides a rigorous asymptotic theory that does not rely on assumptions of smoothness
- In particular, gives generalizations of AIC and BIC
Why have you never hear of it?

- Few people know both algebraic geometry and statistics
- Main results are expressed in the vocabulary of machine learning
- Some practical applications of the theory require further work
Singular models

• A singularity in a statistical model is a point where the dimensionality of the parameter space collapses (e.g. the Fisher information matrix is not of full rank).
• A singular model contains singularities in the parameter space
• Singular models are the rule, not the exception, in hierarchical models
  • Normal mixtures
  • Hidden Markov Models
  • Neural networks
  • Bayes networks
  • ...
Example: 3-component normal mixture

Suppose we observe $Y = (Y_1 \ldots Y_n)$ where

$$p(y_i \mid \mu, \pi) = \sum_{i=1}^{3} \pi_i \phi(y_i - \mu_i)$$

and $\phi$ is the density of a standard normal mixture. The model is parameterized by $\theta = (\mu_1, \mu_2, \mu_3, \pi_1, \pi_2, \pi_3)$ where $\sum_{i=1}^{3} \pi_i = 1$.

Now suppose the true distribution only has 2 components, not 3. We can represent this as

1. $\mu_i = \mu_j$ for any $i \neq j$. Then $\pi_i, \pi_j$ are determined only up to $\pi_i + \pi_j$.
2. $\pi_i = 0$ for any $i$. Then $\mu_i$ is completely undetermined.
Illustration of normal mixture model

Set of parameters
knowledge = singularity

Set of probability distributions
Dimensions of a statistical model

There are two distinct quantities that represent the dimensionality of a singular model:

**The learning coefficient** ($\lambda$) shows how fast the posterior distribution shrinks with increasing sample size.

**The singular fluctuation** ($\nu$) shows how strongly the posterior distribution fluctuates.

Both are *birational invariants* in algebraic geometry.

In regular models

$$\lambda = \mu = d/2$$

where $d$ is the dimension of the parameter space.
Training and generalization errors

- Machine learning distinguishes between:
  - **Training error**: The model fitting criterion applied to the same data set used for estimation
  - **Generalization error**: The model fitting criterion applied to a new data set

- Model choice should be based on the generalization error
- “Big data” problems allow us to split the data into training and validation samples
- For “small data” problems we use full data for estimation
  - Add a *complexity penalty* to the training error to approximate the generalization error (AIC, DIC)
Widely Applicable Information Criteria (WAIC)

\[
\begin{align*}
\text{WAIC}_1 &= \frac{1}{n} \left\{ - \sum_i \log E_{\theta|Y} \{ p (Y_i | \theta) \} + 2\nu \right\} \\
\text{WAIC}_2 &= \frac{1}{n} \left\{ -E_{\theta|Y} \left\{ \sum_i \log p (Y_i | \theta) \right\} + 2\nu \right\}
\end{align*}
\]

where

\[
2\nu \approx \sum_i \text{Var}_{\theta|Y} \{ \log p (Y_i | \theta) \}
\]

Similar, but not identical to Gelman’s approximation to the effective number of parameters \(p_D\) used by R2WinBUGS.

\[
p_D = 2 \text{Var}_{\theta|Y} \left\{ \sum_i \log p (Y_i | \theta) \right\}
\]
Widely Applicable Information Criteria (WAIC)

\begin{align*}
\text{WAIC}_1 &= \frac{1}{n} \left\{ - \sum_i \log E_{\theta \mid Y} \{ p (Y_i \mid \theta) \} + 2\nu \right\} \\
\text{WAIC}_2 &= \frac{1}{n} \left\{ - E_{\theta \mid Y} \left\{ \sum_i \log p (Y_i \mid \theta) \right\} + 2\nu \right\}
\end{align*}

where

\[ 2\nu \approx \sum_i \text{Var}_{\theta \mid Y} \{ \log p (Y_i \mid \theta) \} \]

Similar, but not identical to Gelman’s approximation to the effective number of parameters \( p_D \) used by R2WinBUGS.

\[ p_D = 2 \text{Var}_{\theta \mid Y} \left\{ \sum_i \log p (Y_i \mid \theta) \right\} \]
WAIC vs DIC

- WAIC is an asymptotically correct approximation to the generalization error for singular and non-singular models.
- WAIC is valid even when the model is not true (i.e. $p(Y | \theta)$ is not the data-generating distribution for any $\theta$)
- DIC is derived for under assumptions of asymptotic normality of the posterior distribution of $\theta$, so cannot be applied to singular models.
- DIC is derived under an explicit “good model” assumption that the data generating distribution can be well approximated by $p(Y | \theta)$ for some $\theta$. 
Bayesian Information Criterion

The asymptotic form of the marginal likelihood is

\[
\log p(Y) = \sum_{i}^{n} \log p(Y_i | \hat{\theta}) - \lambda \log(n) + (m - 1) \log \log(n)
\]

- In regular models \( m = 1, \lambda = d/2 \) and we recover Schwarz’s BIC.
- In singular models \( m, \lambda \) depend on the true parameter value. (circular reasoning problem when used for model choice).
- Calculation of \( \lambda \) is hard. Only two model classes have been completely characterized
  1. Reduced rank regression
  2. One-dimensional finite normal mixture models