Bayesian statistics, Occam razor, and model-independent learning of continuous probability densities

Ilya Nemenman
ITP, UCSB

Joint work with:
William Bialek, Princeton University
Bayesian statistics . . .

This is how it compares to other ultimate answers:

Ilya Nemenman, CIMS/NYU Seminar, August 26, 2003
Bayesian statistics . . .

. . . claims that the the answer to the Great Question of Life, The Universe and Everything is not 42, but

This is how it compares to other ultimate answers:

Ilya Nemenman, CIMS/NYU Seminar, August 26, 2003
Bayesian statistics . . .

. . . claims that the answer to the Great Question of Life, The Universe and Everything is not 42, but

\[ P(\text{Model}|\text{Data}) = \frac{P(\text{Data}|\text{Model}) \cdot P(\text{Model})}{P(\text{Data})} \]

This is how it compares to other ultimate answers:

Ilya Nemenman, CIMS/NYU Seminar, August 26, 2003
Bayesian statistics . . .

. . . claims that the answer to the Great Question of Life, The Universe and Everything is not 42, but

\[ P(\text{Model} | \text{Data}) = \frac{P(\text{Data} | \text{Model}) \cdot P(\text{Model})}{P(\text{Data})} \]

This is how it compares to other ultimate answers:

Ilya Nemenman, CIMS/NYU Seminar, August 26, 2003
<table>
<thead>
<tr>
<th>Feature</th>
<th>Bayes</th>
<th>Best others</th>
</tr>
</thead>
<tbody>
<tr>
<td>consistency</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>convergence rates</td>
<td>optimal</td>
<td>optimal</td>
</tr>
<tr>
<td>model selection</td>
<td>?</td>
<td>✓ (disagreement remains)</td>
</tr>
<tr>
<td>use of prior knowledge</td>
<td>✓</td>
<td>?</td>
</tr>
</tbody>
</table>
Bayes  Best others

model selection ✓ (disagreement remains)
use of prior knowledge ? later
Bayes

Best others

model selection

?  

today
Bayesian model selection for finitely parameterizable distributions
Bayesian model selection for finitely parameterizable distributions

\[ P(x) \]

unknown
Bayesian model selection for finitely parameterizable distributions

\[ P(x) \overset{\text{i.i.d.}}{\longrightarrow} X = \{x_1 \cdots x_N\} \]
Bayesian model selection for finitely parameterizable distributions

\[
P(x) \overset{\text{i.i.d.}}{\longrightarrow} X = \{x_1 \cdots x_N\}
\]

Model family \( A \)

\[Q_A(x|\alpha), \dim \alpha = K_A, \mathcal{P}_A(\alpha), \Pr(A)\]
Bayesian model selection for finitely parameterizable distributions

\[ P(x) \overset{\text{i.i.d.}}{\longrightarrow} X = \{x_1 \cdots x_N\} \]

unknown

Model family \( A \)
- \( Q_A(x|\alpha) \)
- \( \dim \alpha = K_A \)
- \( \mathcal{P}_A(\alpha), \ Pr(A) \)

Model family \( B \)
- \( Q_B(x|\beta) \)
- \( \dim \beta = K_B \)
- \( \mathcal{P}_B(\beta), \ Pr(B) \)
Bayesian model selection for finitely parameterizable distributions

\[ P(x) \xrightarrow{\text{i.i.d.}} X = \{x_1 \cdots x_N\} \]

unknown

\[ A \text{ or } B? \]

- **Model family \(A\):**
  - \(Q_A(x|\alpha)\)
  - \(\text{dim}\alpha = K_A\)
  - \(P_A(\alpha), Pr(A)\)

- **Model family \(B\):**
  - \(Q_B(x|\beta)\)
  - \(\text{dim}\beta = K_B\)
  - \(P_B(\beta), Pr(B)\)

Ilya Nemenman, CIMS/NYU Seminar, August 26, 2003
Solution

Find the model with maximum posterior probability!
Solution

Find the model with maximum posterior probability!

For example, for model $A$:

$$P(A|X) = \frac{P(X|A)Pr(A)}{P(X)}$$

$$P(X|A) = \int d\alpha P_A(\alpha) P(X|\alpha) \sim P(X|\alpha_{ML}) \parallel \delta \alpha_{ML} \parallel$$
Solution

Find the model with maximum posterior probability!

For example, for model $A$:

$$P(A|X) = \frac{P(X|A)Pr(A)}{P(X)}$$

$$P(X|A) = \int d\alpha \mathcal{P}_A(\alpha) P(X|\alpha) \sim P(X|\alpha_{ML}) ||\delta\alpha_{ML}||$$

For large $K_A$, $\delta\alpha_{ML}$ (region of “good” $\alpha$) decreases.
More complicated models are penalized!
Solution

Find the model with maximum posterior probability!

For example, for model $A$:

$$P(A|X) = \frac{P(X|A) Pr(A)}{P(X)}$$

$$P(X|A) = \int d\alpha P_A(\alpha) P(X|\alpha) \sim P(X|\alpha_{ML}) ||\delta\alpha_{ML}||$$

For large $K_A$, $\delta\alpha_{ML}$ (region of “good” $\alpha$) decreases. 
More complicated models are penalized!
(See: Bayes factors, Occam factors; Jaynes 1968, 1979)
Large $N$ expansion

Saddle point (large $N$) expansion is almost always valid.

Large $N$ expansion

Saddle point (large $N$) expansion is almost always valid.

\[
\log P(A|X) \rightarrow \sum_i \log Q_A(x_i|\alpha_{ML})
\]

Saddle point (large $N$) expansion is almost always valid.

$$\log P(A|X) \rightarrow \sum_i \log Q_A(x_i|\alpha_{\text{ML}})$$

$$- \frac{K_A}{2} \log N - \log \det \partial^2 \alpha_{\text{ML}} \sum_i \log \frac{Q(x_i|\alpha_{\text{ML}})}{N}$$

Large $N$ expansion

Saddle point (large $N$) expansion is almost always valid.

$$\log P(A|X) \rightarrow \sum_i \log Q_A(x_i|\alpha_{ML})$$

$$- \frac{K_A}{2} \log N - \log \det \partial^2 \alpha_{ML} \sum_i \log Q(x_i|\alpha_{ML})$$

$$+ \log \mathcal{P}(\alpha_{ML}) + o(N^0)$$

Large $N$ expansion

Saddle point (large $N$) expansion is almost always valid.

$$\log P(A|X) \to \sum_i \log Q_A(x_i|\alpha_{ML})$$

$$- \frac{K_A}{2} \log N - \log \det \partial_2^2 \alpha_{ML} \sum_i \log Q(x_i|\alpha_{ML}) \frac{N}{N}$$

$$+ \log \mathcal{P}(\alpha_{ML}) + o(N^0)$$

Large $N$ expansion

Saddle point (large $N$) expansion is almost always valid.

$$
\log P(A|X) \rightarrow \sum_i \log Q_A(x_i|\alpha_{ML})
$$

goodness of fit

$$
- \frac{K_A}{2} \log N - \log \det \partial^2 \alpha_{ML} \frac{\sum_i \log Q(x_i|\alpha_{ML})}{N}
$$

generalization error, fluctuations, complexity; weak dependence on priors

$$
+ \log P(\alpha_{ML}) + o(N^0)
$$

Large $N$ expansion

Saddle point (large $N$) expansion is almost always valid.

$$
\log P(A|X) \to \sum_i \log Q_A(x_i|\alpha_{ML})
$$

- $\frac{K_A}{2} \log N - \log \det \partial^2 \alpha_{ML}$
- $\sum_i \log Q(x_i|\alpha_{ML})$/$N$

generalization error, fluctuations, complexity; weak dependence on priors

$$
+ \log P(\alpha_{ML}) + o(N^0)
$$

Conclusions
Conclusions

- Bayesian inference penalizes for complexity (large $K$)
Conclusions

- Bayesian inference penalizes for complexity (large $K$)
- Fight between the goodness of fit and the complexity selects an optimal model family.
Conclusions

- Bayesian inference penalizes for complexity (large $K$)
- Fight between the goodness of fit and the complexity selects an optimal model family.
- This is a Bayesian analogue of the MDL principle.
Conclusions

- Bayesian inference penalizes for complexity (large $K$)
- Fight between the goodness of fit and the complexity selects an optimal model family.
- This is a Bayesian analogue of the MDL principle.

Does this generalize to infinite-dimensional models?
Estimating density
Estimating density

Standard setting (solving IE)     Fisher–Wald setting (minimizing risk)
Estimating density

Standard setting (solving IE)
\[ F(t) = \int_{-\infty}^{t} Q(x) \, dx \]

Fisher–Wald setting (minimizing risk)
\[ R[Q] = -\int_{-\infty}^{+\infty} \log Q(x) \, dF(x) \]
Estimating density

Standard setting (solving IE)
\[ F(t) = \int_{-\infty}^{t} Q(x) \, dx \]
\[ \frac{1}{N} \sum x_i \Theta(x_i - t) = \int_{-\infty}^{t} Q(x) \, dx \]

Fisher–Wald setting (minimizing risk)
\[ R[Q] = -\int_{-\infty}^{+\infty} \log Q(x) \, dF(x) \]
\[ R_{\text{emp}}[Q] = -\sum x_i \log Q(x_i) \]
Estimating density

Standard setting
(solving IE)
\[ F(t) = \int_{-\infty}^{t} Q(x) dx \]
\[ \frac{1}{N} \sum_{x_i} \Theta(x_i - t) = \int_{-\infty}^{t} Q(x) dx \]

Fisher–Wald setting
(minimizing risk)
\[ R[Q] = -\int_{-\infty}^{+\infty} \log Q(x) dF(x) \]
\[ R_{\text{emp}}[Q] = -\sum x_i \log Q(x_i) \]

Both settings hypersensitive to fluctuations in \( F(t) \).
Smoothing is required.
Bayesian learning for $K \rightarrow \infty$

<table>
<thead>
<tr>
<th>Finite</th>
<th>Infinite</th>
</tr>
</thead>
</table>

Ilya Nemenman, CIMS/NYU Seminar, August 26, 2003
## Bayesian learning for $K \rightarrow \infty$

<table>
<thead>
<tr>
<th>Finite</th>
<th>Infinite</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\phi(x) = - \log \ell_0 Q(x)$</td>
</tr>
</tbody>
</table>
Bayesian learning for $K \to \infty$

<table>
<thead>
<tr>
<th>Finite</th>
<th>Infinite</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\phi(x) = -\log \ell_0 Q(x)$</td>
</tr>
<tr>
<td>$P(\alpha)$</td>
<td>$P[Q] \propto \exp \left[ -\frac{\ell^2 \eta^{-1}}{2} \int dx (\partial_x \phi)^2 \right]$</td>
</tr>
</tbody>
</table>
## Bayesian learning for $K \to \infty$

<table>
<thead>
<tr>
<th>Finite</th>
<th>Infinite</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\phi(x) = -\log \ell_0 Q(x)$</td>
</tr>
<tr>
<td>$\mathcal{P}(\alpha)$</td>
<td>$\mathcal{P}[Q] \propto \exp \left[ -\frac{\ell^2 \eta^{-1}}{2} \int dx (\partial_x^2 \phi)^2 \right]$</td>
</tr>
</tbody>
</table>
## Bayesian learning for $K \rightarrow \infty$

<table>
<thead>
<tr>
<th>Finite</th>
<th>Infinite</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\phi(x) = -\log \ell_0 Q(x)$</td>
</tr>
<tr>
<td>$P(\alpha)$</td>
<td>$P[Q] \propto \exp \left[ -\frac{\ell^{2\eta-1}}{2} \int dx (\partial_x^{2\eta} \phi)^2 \right]$</td>
</tr>
</tbody>
</table>

- Smoothness penalty
- Spline prior of order $2\eta - 1$
## Bayesian learning for $K \to \infty$

<table>
<thead>
<tr>
<th>Finite</th>
<th>Infinite</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\phi(x) = -\log \ell_0 Q(x)$</td>
</tr>
<tr>
<td>$P(\alpha)$</td>
<td>$P[Q] \propto \exp \left[ - \frac{\ell^{2\eta-1}}{2} \int dx (\partial_x^n \phi)^2 \right]$</td>
</tr>
<tr>
<td>${A, K_A}$</td>
<td>${\ell, \eta(?)}$ — index continuum of families</td>
</tr>
</tbody>
</table>

smoothness penalty

spline prior of order $2\eta - 1$
# Bayesian learning for $K \to \infty$

<table>
<thead>
<tr>
<th>Finite</th>
<th>Infinite</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\phi(x) = -\log \ell_0 Q(x)$</td>
</tr>
<tr>
<td>$P(\alpha)$</td>
<td>$P[Q] \propto \exp \left[ -\frac{\ell^{2\eta-1}}{2} \int dx (\partial_x^\eta \phi)^2 \right]$</td>
</tr>
<tr>
<td>${A, K_A}$</td>
<td>smoothness penalty</td>
</tr>
<tr>
<td>$Pr(A)$</td>
<td>spline prior of order $2\eta - 1$</td>
</tr>
<tr>
<td>${\ell, \eta(?)}$</td>
<td>${\ell, \eta(?)} -$ index continuum of families</td>
</tr>
<tr>
<td>$Pr(\ell, \eta(?)$</td>
<td>$Pr(\ell, \eta(?)$</td>
</tr>
</tbody>
</table>

Ilya Nemenman, CIMS/NYU Seminar, August 26, 2003
## Finite vs. Infinite

<table>
<thead>
<tr>
<th>Finite</th>
<th>Infinite</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\phi(x) = -\log \ell_0 Q(x)$</td>
</tr>
<tr>
<td>$P(\alpha)$</td>
<td>$P[Q] \propto \exp \left[ -\frac{\ell^{2\eta-1}}{2} \int dx (\partial_x^n \phi)^2 \right]$</td>
</tr>
<tr>
<td>${A, K_A}$</td>
<td>Smoothness penalty</td>
</tr>
<tr>
<td>$Pr(A)$</td>
<td>Spline prior of order $2\eta - 1$</td>
</tr>
<tr>
<td>${\ell, \eta(?)}$</td>
<td>${\ell, \eta(?)} - \text{index continuum of families}$</td>
</tr>
<tr>
<td>$Pr(\ell, \eta(?)$)</td>
<td></td>
</tr>
</tbody>
</table>

(See: Bialek, Callan, Strong, 1996)
Quantum Field Theory analogy

Fix \( \ell \) and \( \eta \):

\[
\langle Q(x) Q(x_1) \cdots Q(x_N) \rangle^0
\]

Correlation function in a QFT defined by \( \mathcal{P}[Q] \)
Quantum Field Theory analogy

Fix $\ell$ and $\eta$:

$$P[Q|X] = \frac{P(X|Q)P[Q]}{P(X)}$$

$$= \frac{\langle Q(x)Q(x_1)\cdots Q(x_N) \rangle^0}{\langle Q(x_1)\cdots Q(x_N) \rangle^0}$$

Correlation function in a QFT defined by $\mathcal{P}[Q]$
Quantum Field Theory analogy

Fix \( \ell \) and \( \eta \):

\[
P[Q|X] = \frac{P(X|Q)\mathcal{P}[Q]}{P(X)}
\]

\[
\langle Q \rangle = \frac{\int \! dQ \, \mathcal{P}[Q] \, Q(x) \prod_{i=1}^{N} Q(x_i)}{\int \! dQ \, \mathcal{P}[Q] \prod_{i=1}^{N} Q(x_i)}
\]

\[
= \frac{\langle Q(x)Q(x_1)\cdots Q(x_N) \rangle^0}{\langle Q(x_1)\cdots Q(x_N) \rangle^0}
\]

Correlation function in a QFT defined by \( \mathcal{P}[Q] \)
Explicit form of correlation functions

\[ \text{C. F. } \equiv \int [dQ] \mathcal{P}[Q] \prod_{i=1}^{N} Q(x_i) \]

\[ = \int [d\phi] \frac{1}{\ell_0^N} e^{-S[\phi]} \delta \left[ \int dx \frac{1}{\ell_0} e^{-\phi} - 1 \right] \]

\[ S[\phi] \]

\[ \left\{ \begin{array}{l}
\text{action} \\
\text{kinetic term} \\
\text{random potential}
\end{array} \right. \]

\[ = \frac{\ell^{2n-1}}{2} \int dx (\partial_x^n \phi)^2 + \sum_i \phi(x_i) \]
Large $N$ approximation for $\eta = 1$
ML (classical, saddle point) solution dominates
Large $N$ approximation for $\eta = 1$

ML (classical, saddle point) solution dominates

\[ \ell \partial_x^2 \phi_{cl}(x) + \frac{N}{\ell_0} e^{-\phi_{cl}(x)} = \sum_j \delta(x - x_j) \]
Large $N$ approximation for $\eta = 1$

ML (classical, saddle point) solution dominates

$$\ell \partial_x^2 \phi_{cl}(x) + \frac{N}{\ell_0} e^{-\phi_{cl}(x)} = \sum_j \delta(x - x_j)$$

changes on scale
$$\delta x \sim \sqrt{\ell/N P(x)}$$

converges to
$$- \log \ell_0 P(x)$$
Large $N$ approximation for $\eta = 1$

ML (classical, saddle point) solution dominates

$$\ell \frac{d^2}{dx^2} \phi_{cl}(x) + \frac{N}{\ell_0} e^{-\phi_{cl}(x)} = \sum_j \delta(x - x_j)$$

changes on scale $\delta x \sim \sqrt{\ell / NP(x)}$

converges to $- \log \ell_0 P(x)$

changes on scale $\delta x \sim \sqrt{\ell / NP(x)}$

Actual distribution
Fit for 1e5 samples
Fit for 1e3 samples
Fit for 1e1 samples
Large $N$ approximation for $\eta = 1$, continued

Van Vleck calculation of functional determinant:
Large $N$ approximation for $\eta = 1$, continued

Van Vleck calculation of functional determinant:

\[
C. F. \approx \left(\frac{1}{\ell_0}\right)^N e^{-S_{\text{eff}}[\phi_{\text{cl}}(x)]}
\]
Large $N$ approximation for $\eta = 1$, continued

Van Vleck calculation of functional determinant:

$$C. F. \approx (1/\ell_0)^N e^{-S_{\text{eff}}[\phi_{\text{cl}}(x)]}$$

$$S_{\text{eff}}[\phi_{\text{cl}}] = \frac{\ell}{2} \int dx (\partial \phi_{\text{cl}})^2 + \sum \phi_{\text{cl}}(x_i)$$

$$+ \frac{1}{2} \sqrt{\frac{N}{\ell \ell_0}} \int dx e^{-\phi_{\text{cl}}(x)/2}$$
Large $N$ approximation for $\eta = 1$, continued

Van Vleck calculation of functional determinant:

\[
C. F. \approx \left( \frac{1}{\ell_0} \right)^N e^{-S_{\text{eff}}[\phi_{\text{cl}}(x)]}
\]

\[
S_{\text{eff}}[\phi_{\text{cl}}] = \frac{\ell}{2} \int dx (\partial \phi_{\text{cl}})^2 + \sum \phi_{\text{cl}}(x_i)
\]

- prior, smoothness
- goodness of fit

\[
+ \frac{1}{2} \sqrt{\frac{N}{\ell \ell_0}} \int dx e^{-\phi_{\text{cl}}(x)/2}
\]

- fluctuations, complexity, error
How do we measure performance?
How do we measure performance?

For $x \in [0, L)$ the *universal* learning curve is

$$\Lambda(N) \rightarrow \langle D_{KL}(P||Q_{cl}) \rangle^0_{\{x_i\}} \sim \sqrt{\frac{L}{\ell N}}$$
How do we measure performance?

For \( x \in [0, L) \) the \textit{universal} learning curve is

\[
\Lambda(N) \to \langle D_{\text{KL}}(P||Q_{\text{cl}}) \rangle_{\{x_i\}}^0 \sim \sqrt[2]{\frac{L}{\ell N}}
\]

For a different \( \eta \):

\[
\Lambda(N) \sim \left(\frac{L}{\ell}\right)^{1/2\eta} N^{1/2\eta - 1}
\]
Learning curves for fixed $\ell, \eta = 1$
Learning curves for fixed $\ell, \eta = 1$

Learner’s assumptions $\mathcal{P}_{\ell,\eta=1}[Q]$
Learning curves for fixed $\ell, \eta = 1$

Learner’s assumptions: $\mathcal{P}_{\ell,\eta=1}[Q]$

Actual target distribution: $\mathcal{P}_{\ell_a,\eta_a}[Q]$
Learning curves for fixed $\ell, \eta = 1$

Learner’s assumptions $\mathcal{P}_{\ell, \eta=1}[Q]$  
Actual target distribution $\mathcal{P}^\prime_{\ell_a, \eta_a}[Q]$  

$\eta = \eta_a$, $\ell = \ell_a$  learning typical cases, $\mathcal{P} = \mathcal{P}^\prime$
Learning curves for fixed $\ell, \eta = 1$

Learner’s assumptions \[ P_{\ell, \eta=1}[Q] \]
Actual target distribution \[ P'_{\ell_a, \eta_a}[Q] \]

$\eta = \eta_a, \ell = \ell_a$ learning typical cases, $P = P'$
$\eta = \eta_a, \ell \neq \ell_a$ marginal outliers of $P$
Learning curves for fixed $\ell, \eta = 1$

Learner’s assumptions
Actual target distribution

$\eta = \eta_a, \ell = \ell_a$  \hspace{1cm} learning typical cases, $\mathcal{P} = \mathcal{P}'$

$\eta = \eta_a, \ell \neq \ell_a$  \hspace{1cm} marginal outliers of $\mathcal{P}$

$\eta > \eta_a$  \hspace{1cm} extremely rough outliers
Learning curves for fixed $\ell$, $\eta = 1$

Learner’s assumptions $\mathcal{P}_{\ell,\eta=1}[Q]$

Actual target distribution $\mathcal{P}'_{\ell_a,\eta_a}[Q]$

- $\eta = \eta_a$, $\ell = \ell_a$  
  learning typical cases, $\mathcal{P} = \mathcal{P}'$
- $\eta = \eta_a$, $\ell \neq \ell_a$  
  marginal outliers of $\mathcal{P}$
- $\eta > \eta_a$  
  extremely rough outliers
- $\eta < \eta_a$  
  extremely smooth outliers
Learning curves for fixed $\ell, \eta = 1$

Learner’s assumptions

Actual target distribution

$\eta = \eta_a, \ell = \ell_a$ learning typical cases, $P = P'$

$\eta = \eta_a, \ell \neq \ell_a$ marginal outliers of $P$

$\eta > \eta_a$ extremely rough outliers

$\eta < \eta_a$ extremely smooth outliers

Note: we must have $\eta > 1/2$ for convergence of the integrals.
Learning typical cases

\[ \Lambda_N = 0.05, \text{ data and best fit} \]

\[ \Lambda_N = 0.2, \text{ data and best fit} \]

\[ \Lambda_N = 0.4, \text{ data and best fit} \]
\[
\ell = 0.4, \quad \Lambda = (0.54 \pm 0.07)N^{-0.483 \pm 0.014} \\
\ell = 0.2, \quad \Lambda = (0.83 \pm 0.08)N^{-0.493 \pm 0.09} \\
\ell = 0.05, \quad \Lambda = (1.64 \pm 0.16)N^{-0.507 \pm 0.09}
\]
Learning marginal outliers

\[ \ell_a = 0.4, \quad \Lambda = (0.56 \pm 0.08)N^{-0.477\pm0.015} \]

\[ \ell_a = 0.05, \quad \Lambda = (1.90 \pm 0.16)N^{-0.502\pm0.008} \]
Learning at $\ell = 0.2$. 
Learning strong outliers

\[ \eta_a = 2, \ell_a = 0.1, \quad \Lambda = (0.40 \pm 0.05) N^{-0.493 \pm 0.013} \]

\[ \eta_a = 0.8, \ell_a = 0.1, \quad \Lambda = (1.06 \pm 0.08) N^{-0.355 \pm 0.008} \]
\[ \ell = 0.1 \text{ for } \eta_a = 0 \text{ and } \ell = 0.2 \text{ otherwise} \]
Conclusions for fixed $\eta$ and $\ell$

Ilya Nemenman, CIMS/NYU Seminar, August 26, 2003
Conclusions for fixed \( \eta \) and \( \ell \)

- No overfits!
Conclusions for fixed $\eta$ and $\ell$

- No overfits!
- but suboptimal performance for learning outliers
Smoothness scale selection
Smoothness scale selection

Allow a prior over $\ell$, but keep $\eta = 1$

$$C. F. \rightarrow \langle C. F. \rangle_\ell$$
Smoothness scale selection

Allow a prior over \( \ell \), but keep \( \eta = 1 \)

\[
\text{C. F. } \rightarrow \langle \text{C. F.} \rangle_\ell = \int d\ell \ Pr(\ell) \ e^{-S_{\text{eff}}[\phi_{\text{cl}}(\phi,\ell)]}
\]
Smoothness scale selection

Allow a prior over $\ell$, but keep $\eta = 1$

\[
\text{C. F.} \rightarrow \langle \text{C. F.} \rangle_\ell = \int d\ell \ Pr(\ell) \ e^{-S_{\text{eff}}[\phi_{\text{cl}}(\phi,\ell)]}
\]

\[
S_{\text{eff}}[\phi_{\text{cl}}] = \underline{\text{smoothing}} + \underline{\text{data}} + \underline{\text{fluctuations}}
\]
Smoothness scale selection

Allow a prior over $\ell$, but keep $\eta = 1$

\[ \text{C. F.} \rightarrow \langle \text{C. F.} \rangle_\ell = \int d\ell \Pr(\ell) \, e^{-S_{\text{eff}}[\phi_{\text{cl}}(\phi, \ell)]} \]

\[ S_{\text{eff}}[\phi_{\text{cl}}] = \text{smoothing} + \text{data} + \text{fluctuations} \]

- smoothing grows with $\ell$
- data grows with $1/\ell$
Smoothness scale selection

Allow a prior over \( \ell \), but keep \( \eta = 1 \)

\[
C. \ F. \rightarrow \langle C. \ F. \rangle_\ell = \int d\ell \ Pr(\ell) \ e^{-S_{\text{eff}}[\phi_{\text{cl}}(\phi, \ell)]}
\]

\[
S_{\text{eff}}[\phi_{\text{cl}}] = \text{smoothing} + \text{data} + \text{fluctuations}
\]

- grows with \( \ell \)
- grows with \( 1/\ell \)

Some \( \ell^* \) always dominates the C. F. and \( \langle Q \rangle \)!
Calculations: What is \( \ell^* \) for \( \eta_a \) and \( \ell_a \)?

Averaging over \( \ell \) and allowing \( \ell^* = \ell^*(N) \) deals with
Calculations: What is $\ell^*$ for $\eta_a$ and $\ell_a$?

If $\eta = \eta_a$, then $\ell^* = \ell_a$.

Averaging over $\ell$ and allowing $\ell^* = \ell^*(N)$ deals with...
Calculations: What is $\ell^*$ for $\eta_a$ and $\ell_a$?

If $\eta = \eta_a$, then $\ell^* = \ell_a$. Otherwise:

| $0.5 < \eta_a \leq 1.5$ | $1.5 < \eta_a$ |

Averaging over $\ell$ and allowing $\ell^* = \ell^*(N)$ deals with
Calculations: What is $\ell^*$ for $\eta_a$ and $\ell_a$?

If $\eta = \eta_a$, then $\ell^* = \ell_a$. Otherwise:

<table>
<thead>
<tr>
<th>$0.5 &lt; \eta_a \leq 1.5$</th>
<th>$1.5 &lt; \eta_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>data $&gt; \text{smoothing}$</td>
<td>smoothing $&gt; \text{data}$</td>
</tr>
</tbody>
</table>

Averaging over $\ell$ and allowing $\ell^* = \ell^*(N)$ deals with
Calculations: What is $\ell^*$ for $\eta_a$ and $\ell_a$?

If $\eta = \eta_a$, then $\ell^* = \ell_a$. Otherwise:

<table>
<thead>
<tr>
<th>$0.5 &lt; \eta_a \leq 1.5$</th>
<th>$1.5 &lt; \eta_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>data $&gt;$ smoothing</td>
<td>smoothing $&gt;$ data</td>
</tr>
<tr>
<td>$\ell^* \sim N(\eta_a^{-1})/\eta_a$</td>
<td>$\ell^* \sim N^{1/3}$</td>
</tr>
</tbody>
</table>

Averaging over $\ell$ and allowing $\ell^* = \ell^*(N)$ deals with
Calculations: What is $\ell^*$ for $\eta_a$ and $\ell_a$?

If $\eta = \eta_a$, then $\ell^* = \ell_a$. Otherwise:

\[
\begin{array}{|c|c|}
\hline
0.5 < \eta_a \leq 1.5 & 1.5 < \eta_a \\
\hline
\text{data > smoothing} & \text{smoothing > data} \\
\hline
\ell^* \sim N^{(\eta_a - 1)/\eta_a} & \ell^* \sim N^{1/3} \\
\Lambda \sim N^{1/2\eta_a - 1} & \Lambda \sim N^{-2/3} \\
\hline
\end{array}
\]

Averaging over $\ell$ and allowing $\ell^* = \ell^*(N)$ deals with
Calculations: What is $\ell^*$ for $\eta_a$ and $\ell_a$?

If $\eta = \eta_a$, then $\ell^* = \ell_a$. Otherwise:

<table>
<thead>
<tr>
<th>$0.5 &lt; \eta_a \leq 1.5$</th>
<th>$1.5 &lt; \eta_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>data $&gt;$ smoothing</td>
<td>smoothing $&gt;$ data</td>
</tr>
<tr>
<td>$\ell^* \sim N^{(\eta_a-1)/\eta_a}$</td>
<td>$\ell^* \sim N^{1/3}$</td>
</tr>
<tr>
<td>$\Lambda \sim N^{1/2\eta_a-1}$</td>
<td>$\Lambda \sim N^{-2/3}$</td>
</tr>
<tr>
<td>best possible performance</td>
<td>better, but not best performance</td>
</tr>
</tbody>
</table>

Averaging over $\ell$ and allowing $\ell^* = \ell^*(N)$ deals with...
qualitatively wrong smoothness $\eta_a \neq 1!$
Numerics: What is $\ell^*$ for $\eta_a$ and $\ell_a$?
Numerics: What is $\ell^*$ for $\eta_a$ and $\ell_a$?

Note: just single runs shown.

Ilya Nemenman, CIMS/NYU Seminar, August 26, 2003
Numerics: What is $\ell^*$ for $\eta_a$ and $\ell_a$?

Note: just single runs shown.

Ilya Nemenman, CIMS/NYU Seminar, August 26, 2003
Numerics: What is $\ell^*$ for $\eta_a$ and $\ell_a$?

Note: just single runs shown.

Ilya Nemenman, CIMS/NYU Seminar, August 26, 2003
Approaching model–independent optimal inference!
Analogies
Analogies

• choosing $\ell^*$ corresponds to selection of a structure element with $d_{VC} = \sqrt{NL/\ell^*}$ in Vapnik’s SRM theory
Analogies

• choosing $\ell^*$ corresponds to selection of a structure element with $d_{VC} = \sqrt{NL/\ell^*}$ in Vapnik’s SRM theory

• maximizing $P$ over model families ($\ell$’s) asymptotically corresponds to searching for MDL
Analogies

• choosing $\ell^*$ corresponds to selection of a structure element with $d_{VC} = \sqrt{NL/\ell^*}$ in Vapnik’s SRM theory

• maximizing $P$ over model families ($\ell$’s) asymptotically corresponds to searching for MDL

• a lot in common with the Gaussian Processes theory; however normalization constraint is important
Summary

Bayesian smoothness (model) selection works for nonparametric spline priors!
Open questions
Open questions

- constant factor or constant summand?
Open questions

- constant factor or constant summand?
- what to do with $\eta_a > 1.5$?
Open questions

• constant factor or constant summand?
• what to do with $\eta_a > 1.5$?
• reparameterization invariance
Open questions

- constant factor or constant summand?
- what to do with $\eta_a > 1.5$?
- reparameterization invariance
- information theoretic meaningful priors
Open questions

- constant factor or constant summand?
- what to do with $\eta_a > 1.5$?
- reparameterization invariance
- information theoretic meaningful priors
- higher dimensions
Open questions

- constant factor or constant summand?
- what to do with $\eta_a > 1.5$?
- reparameterization invariance
- information theoretic meaningful priors
- higher dimensions
- smooth transition from $K = \text{const}$ to $K \to \infty$
Open questions

- constant factor or constant summand?
- what to do with $\eta_a > 1.5$?
- reparameterization invariance
- information theoretic meaningful priors
- higher dimensions
- smooth transition from $K = \text{const}$ to $K \to \infty$
- which classes of priors are allowed?
Open questions

- constant factor or constant summand?
- what to do with $\eta_a > 1.5$?
- reparameterization invariance
- information theoretic meaningful priors
- higher dimensions
- smooth transition from $K = \text{const}$ to $K \to \infty$
- which classes of priors are allowed?

There is hope that all of this problems are resolvable in a single formulation.