Alternative approaches to Bayesian nonparametrics

Peter Hoff

Statistics, Biostatistics and the CSSS
University of Washington
Challenges for NP Bayes

Adjusted nonparametric priors

Marginal likelihoods

Nonparametric misspecified models
Flexible models for complex systems:

- univariate and multivariate density estimation;
- complicated cluster models;
- hierarchical models;
- nonparametric regression.

NP Bayes provides

- posterior distributions for arbitrary functionals;
- asymptotic consistency;
- and much more.
NP Bayes: A great success

Flexible models for complex systems:

• univariate and multivariate density estimation;
• complicated cluster models;
• hierarchical models;
• nonparametric regression.

NP Bayes provides

• posterior distributions for arbitrary functionals;
• asymptotic consistency;
• and much more.
Is NP Bayes

- Bayesian?
- nonparametric?
What is NP Bayes?

Is NP Bayes

- Bayesian?
- nonparametric?
Is it Bayesian?

Bayesian inference is the change from prior to posterior information:

- the prior information could be your information;
- it can be someone else’s;
- it can even be “really small.”

It should at least approximate information that someone could possibly have.

Otherwise

\[ p(\theta|y) \neq \text{posterior information} \]

because

\[ p(\theta) \neq \text{prior information} \]
Is it Bayesian?

Bayesian inference is the change from prior to posterior information:

- the prior information could be your information;
- it can be someone else’s;
- it can even be “really small.”

It should at least approximate information that someone could possibly have.

Otherwise

\[ p(\theta|y) \neq \text{posterior information} \]

because

\[ p(\theta) \neq \text{prior information} \]
Bayesian inference is the \textbf{change} from prior to posterior information:

- the prior information could be your information;
- it can be someone else’s;
- it can even be “really small.”

It should at least approximate information that someone could possibly have.

Otherwise

$$p(\theta|y) \neq \text{posterior information}$$

because

$$p(\theta) \neq \text{prior information}$$
Bayesian inference is the change from prior to posterior information:

- the prior information could be your information;
- it can be someone else’s;
- it can even be “really small.”

It should at least approximate information that someone could possibly have.

Otherwise

\[ p(\theta|y) \neq \text{posterior information} \]

because

\[ p(\theta) \neq \text{prior information} \]
Is it nonparametric?

If “nonparametric” means no parameters, then clearly not.

If “nonparametric” means consistency for any population, then yes (usually):

- DPM: $p(y|q) = \int f(y|\theta)q(d\theta) \Rightarrow p(y|q) \in \mathcal{H}\{f(y|\theta), \theta \in \Theta\}$
- Examples of inconsistency

A “nonparametric mixture model”?

- “With the Pólya urn, we don’t have to specify the number of clusters”
- Pólya urn is a 1 parameter model for the partition and number of clusters.

Concerns about small sample behavior:

- When is a simpler (parametric model) favored?
- How far off are predictive distributions?
Is it nonparametric?

If “nonparametric” means no parameters, then clearly not.

If “nonparametric” means consistency for any population, then yes (usually):

- DPM: \( p(y|q) = \int f(y|\theta)q(d\theta) \Rightarrow p(y|q) \in \mathcal{H}\{f(y|\theta), \theta \in \Theta\} \)

- Examples of inconsistency

A “nonparametric mixture model”?

- “With the Pólya urn, we don’t have to specify the number of clusters”
- Pólya urn is a 1 parameter model for the partition and number of clusters.

Concerns about small sample behavior:

- When is a simpler (parametric model) favored?
- How far off are predictive distributions?
Is it nonparametric?

If “nonparametric” means no parameters, then clearly not.

If “nonparametric” means consistency for any population, then yes (usually):

- DPM: \( p(y|q) = \int f(y|\theta)q(d\theta) \Rightarrow p(y|q) \in \mathcal{H}\{f(y|\theta), \theta \in \Theta\} \)
- Examples of inconsistency

A “nonparametric mixture model”?

- “With the Pólya urn, we don’t have to specify the number of clusters”
- Pólya urn is a 1 parameter model for the partition and number of clusters.

Concerns about small sample behavior:

- When is a simpler (parametric model) favored?
- How far off are predictive distributions?
Is it nonparametric?

If “nonparametric” means no parameters, then clearly not.

If “nonparametric” means consistency for any population, then yes (usually):
- DPM: \( p(y|q) = \int f(y|\theta)q(d\theta) \Rightarrow p(y|q) \in \mathcal{H}\{f(y|\theta), \theta \in \Theta\} \)
- Examples of inconsistency

A “nonparametric mixture model”?
- “With the Pólya urn, we don’t have to specify the number of clusters”
- Pólya urn is a 1 parameter model for the partition and number of clusters.

Concerns about small sample behavior:
- When is a simpler (parametric model) favored?
- How far off are predictive distributions?
Is it nonparametric?

If “nonparametric” means no parameters, then clearly not.

If “nonparametric” means consistency for any population, then yes (usually):
- DPM: \( p(y|q) = \int f(y|\theta)q(d\theta) \Rightarrow p(y|q) \in \mathcal{H}\{f(y|\theta), \theta \in \Theta\} \)
- Examples of inconsistency

A “nonparametric mixture model”?
- “With the Pólya urn, we don’t have to specify the number of clusters”
- Pólya urn is a 1 parameter model for the partition and number of clusters.

Concerns about small sample behavior:
- When is a simpler (parametric model) favored?
- How far off are predictive distributions?
From Xinyi Xu’s JSM talk

(Xu, MacEachern, Lu, Xu, work in progress)

\[ y_1, \ldots, y_n \sim p_0 = \text{iid skew-normal}(0, 1.5, 2.5) \]

Incorrect parametric model:

\[ p_0(y) \in \{ \text{dnorm}(y, \theta, \sigma^2) : \theta \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+ \} \]

“Correct” nonparametric model:

\[
p_0(y) = \int \text{dnorm}(y, \theta, \sigma^2)q(d\theta)
\]

\[ q \sim DP(\alpha = 2, q_0 = \text{dnorm}(\theta, \mu, \tau^2)) \]
Sequential prediction

\[(Xu, MacEachern, Lu, Xu, work in progress)\]

\[
\log p(y_{n+1}|y_n, \ldots, y_1)
\]
Sequential prediction

\( (Xu, MacEachern, Lu, Xu, \text{ work in progress}) \)

\[
\log p(y_{n+1}|y_n, \ldots, y_1)
\]
Model comparison

(Xu, MacEachern, Lu, Xu, work in progress)

\[
\log \frac{p(y_1, \ldots, y_n | M_0)}{p(y_1, \ldots, y_n | M_0)} = \sum_{i=1}^{n-1} \log \frac{p(y_{i+1} | y_i, \ldots, y_1 | M_0)}{p(y_{i+1} | y_i, \ldots, y_1 | M_1)}
\]
Model comparison

\[ \log \frac{p(y_1, \ldots, y_n | M_0)}{p(y_1, \ldots, y_n | M_0)} = \sum_{i=1}^{n-1} \log \frac{p(y_{i+1} | y_i, \ldots, y_1 | M_0)}{p(y_{i+1} | y_i, \ldots, y_1 | M_1)} \]

Example 1 (Cont.)

Here is the whole story...

\[ \begin{array}{cccccc}
\text{Sample size} & 1 & 21 & 41 & 56 & 71 & 86 & 101 & 126 & 151 & 176 \\
\text{log(Bayes Factor)} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9
\end{array} \]
What is NP Bayes?

\[
\log p(f|y) = \log p(y|f) + \log p(f) + c(y)
\]

\(\log p(f|y)\): An ordering on the parameter space. Estimates and predictions come from this.

\(\log p(y|f)\): A measure of data fidelity. A “good” \(f\) will match the data. This provides the nonparametric consistency.

\(\log p(f)\): A complexity penalty. This is the Bayesian part, without it the best \(f\) is the MLE.

What is the difference between penalized likelihood and Bayesian estimation?

- One uses coordinate descent optimization, the other MCMC?
  - MCMC is unreliable but provides much.
  - Optimization is more reliable but inferentially limited.

- Interpretation?
  - NP Bayes: If \(p(f)\) is prior information, then \(p(f|y)\) is posterior information.
  - Penalized MLE: \(p(f)\) chosen to influence form/sampling characteristics of \(\hat{f}\).
What is NP Bayes?

\[
\log p(f | y) = \log p(y | f) + \log p(f) + c(y)
\]

\( \log p(f | y) \): An ordering on the parameter space. Estimates and predictions come from this.

\( \log p(y | f) \): A measure of data fidelity. A “good” \( f \) will match the data. This provides the nonparametric consistency.

\( \log p(f) \): A complexity penalty. This is the Bayesian part, without it the best \( f \) is the MLE.

What is the difference between penalized likelihood and Bayesian estimation?

- One uses coordinate descent optimization, the other MCMC?
  - MCMC is unreliable but provides much.
  - Optimization is more reliable but inferentially limited.

- Interpretation?
  - NP Bayes: If \( p(f) \) is prior information, then \( p(f | y) \) is posterior information.
  - Penalized MLE: \( p(f) \) chosen to influence form/sampling characteristics of \( \hat{f} \).
What is NP Bayes?

$$\log p(f|y) = \log p(y|f) + \log p(f) + c(y)$$

$$\log p(f|y)$$: An ordering on the parameter space. Estimates and predictions come from this.

$$\log p(y|f)$$: A measure of data fidelity. A “good” $f$ will match the data. This provides the nonparametric consistency.

$$\log p(f)$$: A complexity penalty. This is the Bayesian part, without it the best $f$ is the MLE.

What is the difference between penalized likelihood and Bayesian estimation?

- One uses coordinate descent optimization, the other MCMC?
  - MCMC is unreliable but provides much.
  - Optimization is more reliable but inferentially limited.

- Interpretation?
  - NP Bayes: If $p(f)$ is prior information, then $p(f|y)$ is posterior information.
  - Penalized MLE: $p(f)$ chosen to influence form/sampling characteristics of $\hat{f}$. 
What is NP Bayes?

\[ \log p(f|y) = \log p(y|f) + \log p(f) + c(y) \]

\[ \log p(f|y) \] : An ordering on the parameter space. Estimates and predictions come from this.

\[ \log p(y|f) \] : A measure of data fidelity. A “good” \( f \) will match the data. This provides the nonparametric consistency.

\[ \log p(f) \] : A complexity penalty. This is the Bayesian part, without it the best \( f \) is the MLE.

What is the difference between penalized likelihood and Bayesian estimation?

• One uses coordinate descent optimization, the other MCMC?
  • MCMC is unreliable but provides much.
  • Optimization is more reliable but inferentially limited.

• Interpretation?
  • NP Bayes: If \( p(f) \) is prior information, then \( p(f|y) \) is posterior information.
  • Penalized MLE: \( p(f) \) chosen to influence form/sampling characteristics of \( \hat{f} \).
What is NP Bayes?

\[ \log p(f|y) = \log p(y|f) + \log p(f) + c(y) \]

\( \log p(f|y) \): An ordering on the parameter space. Estimates and predictions come from this.

\( \log p(y|f) \): A measure of data fidelity. A “good” \( f \) will match the data. This provides the nonparametric consistency.

\( \log p(f) \): A complexity penalty. This is the Bayesian part, without it the best \( f \) is the MLE.

What is the difference between penalized likelihood and Bayesian estimation?

- One uses coordinate descent optimization, the other MCMC?
  - MCMC is unreliable but provides much.
  - Optimization is more reliable but inferentially limited.

- Interpretation?
  - NP Bayes: If \( p(f) \) is prior information, then \( p(f|y) \) is posterior information.
  - Penalized MLE: \( p(f) \) chosen to influence form/sampling characteristics of \( \hat{f} \).
NP Bayes and data analysis

Data analysis: \[ y \rightarrow (\text{data analysis algorithm}) \rightarrow t(y) \]

Two measures by which we might evaluate an algorithm include
- reliability: numerical stability of the algorithm
- interpretability: do we understand what it is doing

NP Bayes: \[ y \rightarrow (\text{prior+MCMC}) \rightarrow t(y), \theta_1, \theta_2, \ldots \]

Challenges:
- reliability: MCMC approximation error
- interpretability: prior, MCMC and posterior
Summary of Part 1

Bayesian
- difficult to put meaningful priors on all parameters

Nonparametric
- difficult to estimate all parameters from the data
- difficult to reliably approximate the estimates

Can these issues be addressed?
- Can we make our nonparametric priors more meaningful?
- Can we do NPB inference without priors over nuisance parameters?

Some partly-baked ideas (the rest of this talk)
1. Adjusting nonparametric priors (half-baked)
2. Marginal likelihoods for semiparametric inference (mostly-baked)
3. “Misspecified” models for specific functionals (over-baked)
Summary of Part 1

Bayesian
- difficult to put meaningful priors on all parameters

Nonparametric
- difficult to estimate all parameters from the data
- difficult to reliably approximate the estimates

Can these issues be addressed?
- Can we make our nonparametric priors more meaningful?
- Can we do NPB inference without priors over nuisance parameters?

Some partly-baked ideas (the rest of this talk)
1. Adjusting nonparametric priors (half-baked)
2. Marginal likelihoods for semiparametric inference (mostly-baked)
3. “Misspecified” models for specific functionals (over-baked)
Summary of Part 1

Bayesian
- difficult to put meaningful priors on all parameters

Nonparametric
- difficult to estimate all parameters from the data
- difficult to reliably approximate the estimates

Can these issues be addressed?
- Can we make our nonparametric priors more meaningful?
- Can we do NPB inference without priors over nuisance parameters?

Some partly-baked ideas (the rest of this talk)
1. Adjusting nonparametric priors (half-baked)
2. Marginal likelihoods for semiparametric inference (mostly-baked)
3. “Misspecified” models for specific functionals (over-baked)
Prior conflict

\[ y_1, \ldots, y_n | f \sim_{iid} f \]

\[ f \sim \pi_0(f) \]

The prior \( \pi_0 \) on \( f \) induces priors on all functionals \( \theta = \theta(f) \):

\[ f \sim \pi_0 \Rightarrow \theta \sim p_0(\theta), \text{ where} \]

\[ \Pr(\theta(f) \in A) = \Exp[\delta_A(\theta(f))] = \int_A p_0(\theta) d\theta \]

Suppose we have a functional of interest \( \theta \)
- maybe we have real prior information on \( \theta \),
- maybe we want to compare posteriors under different priors on \( \theta \).

We want: \( p_1(\theta | y) \) under prior \( p_1(\theta) \)
We have: \( p_0(\theta | y) \) under prior \( p_0(\theta) \).

What if \( p_1(\theta) \) and \( p_0(\theta) \) conflict?
Prior conflict

\[ y_1, \ldots, y_n | f \overset{iid}{\sim} f \]
\[ f \sim \pi_0(f) \]

The prior \( \pi_0 \) on \( f \) induces priors on all functionals \( \theta = \theta(f) \):

\[ f \sim \pi_0 \Rightarrow \theta \sim p_0(\theta) \], where

\[ \Pr(\theta(f) \in A) = \exp[\delta_A(\theta(f))] = \int_A p_0(\theta) d\theta \]

Suppose we have a functional of interest \( \theta \)

- maybe we have real prior information on \( \theta \),
- maybe we want to compare posteriors under different priors on \( \theta \).

We want: \( p_1(\theta | y) \) under prior \( p_1(\theta) \)
We have: \( p_0(\theta | y) \) under prior \( p_0(\theta) \).

What if \( p_1(\theta) \) and \( p_0(\theta) \) conflict?
Prior conflict

\[ y_1, \ldots, y_n \mid f \overset{\text{iid}}{\sim} f \]

\[ f \sim \pi_0(f) \]

The prior \( \pi_0 \) on \( f \) induces priors on all functionals \( \theta = \theta(f) \):

\[ f \sim \pi_0 \implies \theta \sim p_0(\theta), \text{ where} \]

\[ \Pr(\theta(f) \in A) = \text{Exp}[\delta_A(\theta(f))] = \int_A p_0(\theta) d\theta \]

Suppose we have a functional of interest \( \theta \)

- maybe we have real prior information on \( \theta \),
- maybe we want to compare posteriors under different priors on \( \theta \).

**We want:** \( p_1(\theta \mid y) \) under prior \( p_1(\theta) \)

**We have:** \( p_0(\theta \mid y) \) under prior \( p_0(\theta) \).

What if \( p_1(\theta) \) and \( p_0(\theta) \) conflict?
Centering nonparametric priors

In normal models, it is possible to choose hyperparameters to induce a particular prior on the population mean:


but in general this will be difficult.
Parametrically adjusted nonparametric priors

Can we combine the nonparametric $\pi_0(f)$ with the parametric $p_1(\theta)$?

$$\pi_0(f) = \pi_0(f|\theta)\pi_0(\theta)$$
$$= \pi_0(f|\theta)p_0(\theta)$$

$$\pi_1(f) = \pi_0(f|\theta)p_1(\theta)$$

More formally,

$$\pi_1(A) \equiv E_{\pi_1}[\delta_A(f)] = \int E_{\pi_0}[\delta_A(f)|\theta]p_1(\theta)d\theta.$$ 

Then

$$\Pr_{\pi_1}(\theta(f) \in A) = \text{Exp}[\delta_A(\theta(f))] = \int_A p_1(\theta)d\theta$$
Can we combine the nonparametric $\pi_0(f)$ with the parametric $p_1(\theta)$?

\[
\pi_0(f) = \pi_0(f|\theta)p_0(\theta) = \pi_0(f|\theta)p_0(\theta)
\]

\[
\pi_1(f) = \pi_0(f|\theta)p_1(\theta)
\]

More formally,

\[
\pi_1(A) \equiv E_{\pi_1}[\delta_A(f)] = \int E_{\pi_0}[\delta_A(f)|\theta]p_1(\theta)d\theta.
\]

Then

\[
\Pr_{\pi_1}(\theta(f) \in A) = \Exp[\delta_A(\theta(f))] = \int_A p_1(\theta)d\theta
\]
Parametrically adjusted nonparametric priors

Can we combine the nonparametric $\pi_0(f)$ with the parametric $p_1(\theta)$?

$$
\pi_0(f) = \pi_0(f|\theta)\pi_0(\theta) = \pi_0(f|\theta)p_0(\theta)
$$

$$
\pi_1(f) = \pi_0(f|\theta)p_1(\theta)
$$

More formally,

$$
\pi_1(A) \equiv E_{\pi_1}[\delta_A(f)] = \int E_{\pi_0}[\delta_A(f)|\theta]p_1(\theta)d\theta.
$$

Then

$$
\Pr_{\pi_1}(\theta(f) \in A) = \Exp[\delta_A(\theta(f))] = \int_A p_1(\theta)d\theta
$$
Parametrically adjusted nonparametric priors

Can we combine the nonparametric $\pi_0(f)$ with the parametric $p_1(\theta)$?

$$
\pi_0(f) = \pi_0(f|\theta)\pi_0(\theta) = \pi_0(f|\theta)p_0(\theta)
$$

$$
\pi_1(f) = \pi_0(f|\theta)p_1(\theta)
$$

More formally,

$$
\pi_1(A) \equiv E_{\pi_1}[\delta_A(f)] = \int E_{\pi_0}[\delta_A(f)|\theta]p_1(\theta)d\theta.
$$

Then

$$
Pr_{\pi_1}(\theta(f) \in A) = \text{Exp}[^A[\delta_A(\theta(f)))] = \int_A p_1(\theta)d\theta
$$
Consistency

The support of this prior can be related to that of $\pi_0$ as follows:

\[
\Pr(f \in A)_{\pi_1} = \int E_{\pi_0}[\delta_A(f)|\theta]p_1(\theta)d\theta
\]

\[
= \int E_{\pi_0}[\delta_A(f)|\theta]\frac{p_1(\theta)}{p_0(\theta)}p_0(\theta)d\theta
\]

\[
= E_{\pi_0}[\delta_A(f)\frac{p_1(\theta)}{p_0(\theta)}]
\]

If $p_1(\theta)/p_0(\theta) > 0$ on $A$, then $\pi_0(A) > 0 \Rightarrow \pi_1(A) > 0$ .

If $p_1(\theta) > 0$ for all $\theta$, then $\text{support}(\pi_1) = \text{support}(\pi_0)$.

Support of prior determines consistency
Consistency

The support of this prior can be related to that of $\pi_0$ as follows:

$$\Pr_{\pi_1}(f \in A) = \int E_{\pi_0} [\delta_A(f) | \theta] p_1(\theta) d\theta$$

$$= \int E_{\pi_0} [\delta_A(f) | \theta] \frac{p_1(\theta)}{p_0(\theta)} p_0(\theta) d\theta$$

$$= E_{\pi_0} [\delta_A(f) \frac{p_1(\theta)}{p_0(\theta)}]$$

If $p_1(\theta)/p_0(\theta) > 0$ on $A$, then $\pi_0(A) > 0 \Rightarrow \pi_1(A) > 0$.

If $p_1(\theta) > 0$ for all $\theta$, then $\text{support}(\pi_1) = \text{support}(\pi_0)$.

Support of prior determines consistency
Consistency

The support of this prior can be related to that of $\pi_0$ as follows:

$$
\Pr(f \in A) = \int E_{\pi_0}[\delta_A(f)|\theta] p_1(\theta) d\theta \tag{1}
$$

$$
= \int E_{\pi_0}[\delta_A(f)|\theta] \frac{p_1(\theta)}{p_0(\theta)} p_0(\theta) d\theta \tag{2}
$$

$$
= E_{\pi_0}[\delta_A(f) \frac{p_1(\theta)}{p_0(\theta)}] \tag{3}
$$

If $\frac{p_1(\theta)}{p_0(\theta)} > 0$ on $A$, then $\pi_0(A) > 0 \Rightarrow \pi_1(A) > 0$.

If $p_1(\theta) > 0$ for all $\theta$, then $\text{support}(\pi_1) = \text{support}(\pi_0)$.

Support of prior determines consistency
NPBayes can be made “more Bayesian” via parametrically adjusted priors

**Applications**:  
- means, variances, quantiles  
- regression coefficients  
- more exotic functionals (KL divergence from densities or models)

**Limitations**:  
- not clear if it is (generally) computationally feasible  
- prior $\pi_1(f|\theta)$ is still “non-Bayesian”
Marginal likelihood

\[ Y \sim p(y|\theta, \psi) \]

- \( \theta \) is the parameter of interest
- \( \psi \) is the nuisance parameter, possibly high dimensional

Suppose we have a statistic \( t() \) such that

\[ p(t(y)|\theta, \psi) = p(t(y)|\theta) \]

Then

\[ p(y|\theta, \psi) = p(t(y), y|\theta, \psi) \]
Marginal likelihood

\[ Y \sim p(y|\theta, \psi) \]

- \( \theta \) is the parameter of interest
- \( \psi \) is the nuisance parameter, possibly high dimensional

Suppose we have a statistic \( t() \) such that

\[ p(t(y)|\theta, \psi) = p(t(y)|\theta) \]

Then

\[
\begin{align*}
p(y|\theta, \psi) &= p(t(y), y|\theta, \psi) \\
&= p(t(y)|\theta, \psi) \times p(y|t(y), \theta, \psi)
\end{align*}
\]
Marginal likelihood

\[ Y \sim p(y|\theta, \psi) \]

- \( \theta \) is the parameter of interest
- \( \psi \) is the nuisance parameter, possibly high dimensional

Suppose we have a statistic \( t() \) such that

\[ p(t(y)|\theta, \psi) = p(t(y)|\theta) \]

Then

\[
p(y|\theta, \psi) = p(t(y), y|\theta, \psi) \\
= p(t(y)|\theta, \psi) \times p(y|t(y), \theta, \psi) \\
= p(t(y)|\theta) \times p(y|t(y), \theta, \psi)
\]

A marginal likelihood estimate of \( \theta \) can be obtained from \( p(t(y)|\theta) \). Specification or estimation of \( \psi \) is not necessary.
Multivariate data

Survey data often yield multivariate data of varied types.

**Hypothetical survey data:** A vector of responses $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,p})$ for each person $i$ in a sample of survey respondents, $i \in \{1, \ldots, n\}$.

- $y_{i,1}$ = income
- $y_{i,2}$ = education level
- $y_{i,3}$ = number of children
- $y_{i,4}$ = age
- $y_{i,5}$ = attitude (Likert scale)

A mix of continuous and discrete ordinal data.
GSS data
Inverse normal model

One possibility would be to transform the data to have normal marginals, then fit a multivariate normal model. This cannot be done for discrete data, but such data can be viewed as a function of normal data.

If $F$ is a distribution there exists a nondecreasing function $g$ such that

1. if $Z \sim \text{normal}(0,1),$
2. and $Y = g(Z),$

then $Y \sim F.$

If $F$ is continuous then $g(z) = F^{-1}(\Phi(z)),$ $g^{-1}$ is a function and $g^{-1}(Y)$ is standard normal. If $F$ is not continuous then $g^{-1}$ maps to a set (this includes probit models, for example).
Multivariate normal copula model

This idea motivates the following “latent variable” model:

\[
(Z_1, \ldots, Z_p) \sim \text{multivariate normal}(0, \Sigma) \\
(Y_1, \ldots, Y_p) = (g_1(Z_1), \ldots, g_p(Z_p))
\]

\(\Sigma\) parameterizes the dependence, \(g_1, \ldots, g_p\) the marginal distributions.

- scale free
- appropriate for discrete and continuous data
- compatible full conditional distributions

Estimation strategies:

- estimation of \(\Sigma\) conditional on plug-in estimates of \(g_1, \ldots, g_p\);
  (procedures for continuous data gives inconsistent results for discrete data)
- joint estimation of \(\Sigma\) and \(g_1, \ldots, g_p\);
  (parametric models of \(g\) too simple, nonparametric too complex)
- marginal likelihood estimation.
  (how would that work?)
This idea motivates the following “latent variable” model:

\[
(Z_1, \ldots, Z_p) \sim \text{multivariate normal}(0, \Sigma)
\]

\[
(Y_1, \ldots, Y_p) = (g_1(Z_1), \ldots, g_p(Z_p))
\]

\(\Sigma\) parameterizes the dependence, \(g_1, \ldots, g_p\) the marginal distributions.

- scale free
- appropriate for discrete and continuous data
- compatible full conditional distributions

**Estimation strategies:**

- estimation of \(\Sigma\) conditional on plug-in estimates of \(g_1, \ldots, g_p\);
  (procedures for continuous data gives inconsistent results for discrete data)

- joint estimation of \(\Sigma\) and \(g_1, \ldots, g_p\);
  (parametric models of \(g\) too simple, nonparametric too complex)

- marginal likelihood estimation.
  (how would that work?)
Multivariate normal copula model

This idea motivates the following “latent variable” model:

\[(Z_1, \ldots, Z_p) \sim \text{multivariate normal}(0, \Sigma)\]
\[(Y_1, \ldots, Y_p) = (g_1(Z_1), \ldots, g_p(Z_p))\]

\(\Sigma\) parameterizes the dependence, \(g_1, \ldots, g_p\) the marginal distributions.

- scale free
- appropriate for discrete and continuous data
- compatible full conditional distributions

Estimation strategies:

- estimation of \(\Sigma\) conditional on plug-in estimates of \(g_1, \ldots, g_p\);
  (procedures for continuous data gives inconsistent results for discrete data)

- joint estimation of \(\Sigma\) and \(g_1, \ldots, g_p\);
  (parametric models of \(g\) too simple, nonparametric too complex)

- marginal likelihood estimation.
  (how would that work?)
This idea motivates the following “latent variable” model:

\[(Z_1, \ldots, Z_p) \sim \text{multivariate normal}(0, \Sigma)\]

\[(Y_1, \ldots, Y_p) = (g_1(Z_1), \ldots, g_p(Z_p))\]

\(\Sigma\) parameterizes the dependence, \(g_1, \ldots, g_p\) the marginal distributions.

- scale free
- appropriate for discrete and continuous data
- compatible full conditional distributions

**Estimation strategies:**

- estimation of \(\Sigma\) conditional on plug-in estimates of \(g_1, \ldots, g_p\);
  (procedures for continuous data gives inconsistent results for discrete data)

- joint estimation of \(\Sigma\) and \(g_1, \ldots, g_p\);
  (parametric models of \(g\) too simple, nonparametric too complex)

- marginal likelihood estimation.
  (how would that work?)
Multivariate normal copula model

This idea motivates the following “latent variable” model:

\[ (Z_1, \ldots, Z_p) \sim \text{multivariate normal}(\mathbf{0}, \Sigma) \]
\[ (Y_1, \ldots, Y_p) = (g_1(Z_1), \ldots, g_p(Z_p)) \]

\(\Sigma\) parameterizes the dependence, \(g_1, \ldots, g_p\) the marginal distributions.

- scale free
- appropriate for discrete and continuous data
- compatible full conditional distributions

Estimation strategies:

- estimation of \(\Sigma\) conditional on plug-in estimates of \(g_1, \ldots, g_p\);
  (procedures for continuous data gives inconsistent results for discrete data)
- joint estimation of \(\Sigma\) and \(g_1, \ldots, g_p\);
  (parametric models of \(g\) too simple, nonparametric too complex)
- marginal likelihood estimation.
  (how would that work?)
Multivariate normal copula model

This idea motivates the following “latent variable” model:

\[ (Z_1, \ldots, Z_p) \sim \text{multivariate normal}(0, \Sigma) \]
\[ (Y_1, \ldots, Y_p) = (g_1(Z_1), \ldots, g_p(Z_p)) \]

\( \Sigma \) parameterizes the dependence, \( g_1, \ldots, g_p \) the marginal distributions.

- scale free
- appropriate for discrete and continuous data
- compatible full conditional distributions

Estimation strategies:

- estimation of \( \Sigma \) conditional on plug-in estimates of \( g_1, \ldots, g_p \);
  (procedures for continuous data gives inconsistent results for discrete data)
- joint estimation of \( \Sigma \) and \( g_1, \ldots, g_p \);
  (parametric models of \( g \) too simple, nonparametric too complex)
- marginal likelihood estimation.
  (how would that work?)
Semiparametric Gaussian copula model:

\[ Z_1, \ldots, Z_n \sim \text{i.i.d. multivariate normal}(0, \Sigma) \]
\[ Y_{i,j} = g_j(Z_{i,j}) \]

- \( \Sigma \) is the parameter of interest
- \( g_1, \ldots, g_p \) are high-dimensional nuisance parameters

For continuous data, let \( r_{i,j} = \text{rank of } y_{i,j} \text{ among } y_{1,j}, \ldots, y_{n,j} \). Then

\[ p(y|\Sigma, g) = p(r, y|\Sigma, g) \]
Semiparametric Gaussian copula model:

\[ Z_1, \ldots, Z_n \sim \text{i.i.d. multivariate normal}(0, \Sigma) \]
\[ Y_{i,j} = g_j(Z_{i,j}) \]

- \( \Sigma \) is the parameter of interest
- \( g_1, \ldots, g_p \) are high-dimensional nuisance parameters

For continuous data, let \( r_{i,j} = \text{rank of } y_{i,j} \text{ among } y_{1,j}, \ldots, y_{n,j} \). Then

\[
p(y|\Sigma, g) = p(r, y|\Sigma, g) = p(r|\Sigma, g) \times p(y|r, \Sigma, g)
\]
Semiparametric Gaussian copula model:

\[ Z_1, \ldots, Z_n \sim \text{i.i.d. multivariate normal}(0, \Sigma) \]

\[ Y_{i,j} = g_j(Z_{i,j}) \]

- \( \Sigma \) is the parameter of interest
- \( g_1, \ldots, g_p \) are high-dimensional nuisance parameters

For continuous data, let \( r_{i,j} = \text{rank of } y_{i,j} \text{ among } y_{1,j}, \ldots, y_{n,j} \). Then

\[
p(y|\Sigma, g) = p(r, y|\Sigma, g) = p(r|\Sigma, g) \times p(y|r, \Sigma, g) = p(r|\Sigma) \times p(y|r, \Sigma, g)
\]

Will this work for discrete data?
Extending the rank likelihood

If $g_j$ is not strictly increasing then

- variable $j$ has atoms,
- $Z_{i_1.j} < Z_{i_2.j} \not\Rightarrow Y_{i_1.j} < Y_{i_2.j}$,
- $p(r|\Sigma, g)$ depends on $g$.

So the rank likelihood depends on $g$.

However, $Y_{i_1.j} < Y_{i_2.j} \Rightarrow Z_{i_1.j} < Z_{i_2.j}$. This means that given $Y = y$ we do know

$$Z \in A(y) = \{z : z_{i_1.j} < z_{i_2.j} \text{ if } y_{i_1.j} < y_{i_2.j}\}$$

We can construct the following marginal likelihood:

$$p(y|\Sigma, g) = p(Z \in A(y), y|\Sigma, g)$$
Extending the rank likelihood

If $g_j$ is not strictly increasing then

- variable $j$ has atoms,
- $Z_{i_1,j} < Z_{i_2,j} \nRightarrow Y_{i_1,j} < Y_{i_2,j}$,
- $p(r|\Sigma, g)$ depends on $g$.

So the rank likelihood depends on $g$.

However, $Y_{i_1,j} < Y_{i_2,j} \Rightarrow Z_{i_1,j} < Z_{i_2,j}$. This means that given $Y = y$ we do know

$$Z \in A(y) = \{z : z_{i_1,j} < z_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j}\}$$

We can construct the following marginal likelihood:

$$p(y|\Sigma, g) = p(Z \in A(y), y|\Sigma, g)$$
Extending the rank likelihood

If $g_j$ is not strictly increasing then

- variable $j$ has atoms,
- $Z_{i_1,j} < Z_{i_2,j} \not\Rightarrow Y_{i_1,j} < Y_{i_2,j},$
- $p(r|\Sigma, g)$ depends on $g.$

So the rank likelihood depends on $g.$

However, $Y_{i_1,j} < Y_{i_2,j} \Rightarrow Z_{i_1,j} < Z_{i_2,j}.$ This means that given $Y = y$ we do know $Z \in A(y) = \{z : z_{i_1,j} < z_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j}\}$

We can construct the following marginal likelihood:

$$p(y|\Sigma, g) = p(Z \in A(y), y|\Sigma, g)$$
Extending the rank likelihood

If $g_j$ is not strictly increasing then

- variable $j$ has atoms,
- $Z_{i_1,j} < Z_{i_2,j} \not\Rightarrow Y_{i_1,j} < Y_{i_2,j}$,
- $p(r|\Sigma, g)$ depends on $g$.

So the rank likelihood depends on $g$.

However, $Y_{i_1,j} < Y_{i_2,j} \Rightarrow Z_{i_1,j} < Z_{i_2,j}$. This means that given $Y = y$ we do know

$$Z \in A(y) = \{z : z_{i_1,j} < z_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j}\}$$

We can construct the following marginal likelihood:

$$p(y|\Sigma, g) = p(Z \in A(y), y|\Sigma, g)$$
Extending the rank likelihood

If $g_j$ is not strictly increasing then

- variable $j$ has atoms,
- $Z_{i_1,j} < Z_{i_2,j} \nRightarrow Y_{i_1,j} < Y_{i_2,j}$,
- $p(r|\Sigma, g)$ depends on $g$.

So the rank likelihood depends on $g$.

However, $Y_{i_1,j} < Y_{i_2,j} \Rightarrow Z_{i_1,j} < Z_{i_2,j}$. This means that given $Y = y$ we do know

$Z \in A(y) = \{z : z_{i_1,j} < z_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j}\}$

We can construct the following marginal likelihood:

$$p(y|\Sigma, g) = p(Z \in A(y), y|\Sigma, g)$$
$$= \Pr(Z \in A(y)|\Sigma, g) \times p(y|Z \in A(y), \Sigma, g)$$
Extending the rank likelihood

If \( g_j \) is not strictly increasing then

- variable \( j \) has atoms,
- \( Z_{i_1,j} < Z_{i_2,j} \not\Rightarrow Y_{i_1,j} < Y_{i_2,j} \),
- \( p(r|\Sigma, g) \) depends on \( g \).

So the rank likelihood depends on \( g \).

However, \( Y_{i_1,j} < Y_{i_2,j} \Rightarrow Z_{i_1,j} < Z_{i_2,j} \). This means that given \( Y = y \) we do know

\[
Z \in A(y) = \{ z : z_{i_1,j} < z_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j} \}
\]

We can construct the following marginal likelihood:

\[
p(y|\Sigma, g) = p(Z \in A(y), y|\Sigma, g) \\
= \Pr(Z \in A(y)|\Sigma, g) \times p(y|Z \in A(y), \Sigma, g) \\
= \Pr(Z \in A(y)|\Sigma) \times p(y|Z \in A(y), \Sigma, g)
\]

\[
\Pr(Z \in A(y)|\Sigma) = \int_{A(y)} \prod_i p(z_i|\Sigma) \, dz_i
\]

If \( g_j \)'s are continuous, then \( \Pr(Z \in A(y)|\Sigma) = \Pr(R = r|\Sigma) \).
Bayesian estimates are easy to obtain.

Given a prior distribution $p(\Sigma)$, we iterate the following steps:

1. for each $i, j$, sample $Z_{i,j} \sim p(Z_{i,j}|\Sigma, Z_{-(i,j)}, Z \in A(y))$,
2. sample $\Sigma \sim p(\Sigma|Z, Z \in A(y)) = p(\Sigma|Z)$.

This generates a Markov chain $\{\Sigma^{(1)}, \Sigma^{(2)}, \ldots\}$ such that

$$\Sigma^{(s)} \overset{d}{\rightarrow} p(\Sigma|Z \in A(y)).$$
The actual R-code

Given \{Z,S\} and \{Ranks,n,p,S0,n0\}:

#### update S

\[
S <- \text{solve}\left(\text{rwish}\left(\text{solve}(S0*n0+t(Z)'*Z),n0+n)\right)\right)
\]

#### update Z

for (j in 1:p) {

\[
Sjc <- S[j,-j]'*\text{solve}(S[-j,-j])
\]
\[
sdj <- \text{sqrt}\left( S[j,j] - S[j,-j]'*\text{solve}(S[-j,-j])'*S[-j,j] \right)
\]
\[
muj <- Z[-j]'*t(Sjc)
\]

for (r in unique(Ranks[,j])){

\[
\text{ir} <- (1:n)[\text{Ranks}[,j]==r & !\text{is.na}(\text{Ranks}[,j])]
\]
\[
\text{lb} <- \text{suppressWarnings}(\text{max}( Z[ \text{Ranks}[,j]==r-1,j],\text{na.rm=TRUE} ))
\]
\[
\text{ub} <- \text{suppressWarnings}(\text{min}( Z[ \text{Ranks}[,j]==r+1,j],\text{na.rm=TRUE} ))
\]
\[
Z[\text{ir},j] <- \text{qnorm}(\text{runif}(\text{length}(\text{ir}),
\text{pnorm}(\text{lb},\text{muj}[\text{ir}],\text{sdj}),\text{pnorm}(\text{ub},\text{muj}[\text{ir}],\text{sdj})),\text{muj}[\text{ir}],\text{sdj})
\]

\[
\text{ir} <- (1:n)[\text{is.na}(\text{Ranks}[,j])]
\]
\[
Z[\text{ir},j] <- \text{rnorm}(\text{length}(\text{ir}),\text{muj}[\text{ir}],\text{sdj})
\]

}
GSS Example

Data on 1002 male respondents to the 1994 GSS.

INC : income of respondent
DEG : highest degree obtained
CHILD : number of children
PINC : income category of parents
PDEG : maximum of mother’s and father’s highest degree
PCHILD : number of siblings plus one
AGE : age in years

Using MCMC integration, we estimate

\[ \Sigma, \] the correlation matrix, and

\[ \Sigma_{[j,-j]} \Sigma_{[-j,-j]}^{-1}, \] the regression coefficients.
MCMC diagnostics

The image shows a set of time series plots labeled as $C_{ij}$, with the x-axis representing scan numbers ranging from 0 to 25000 and the y-axis showing values ranging from -0.2 to 0.6. The plots appear to be fluctuating, indicating some form of diagnostic analysis in the context of MCMC. The specific content or analysis method is not detailed within the image.
Correlations and regressions
Challenges for NP Bayes Adjusted nonparametric priors Marginal likelihoods Nonparametric misspecified models

Correlations and regressions

INF

CHILD

DEG

PDEC

PINC

CHILD

PCHILD
The general transformation model

\[ Z \sim p(z|\theta) \]
The general transformation model

$Z \sim p(z|\theta) \quad Y = g(Z)$
The general transformation model

\[ Z \sim p(z|\theta) \quad Y = g(Z) \quad Z \in A(Y) \]
Suppose we have a set valued function $A() : \mathcal{Y} \rightarrow \sigma(\mathcal{Z})$ such that

$$g^{-1}(y) \subset A(y) \quad \forall y, g,$$

or equivalently,

$$z \in A(g(z)) \quad \forall z, g,$$

Then $\Pr(Z \in A(Y) | \theta, g) = 1$, so

$$\Pr(Y = y | \theta, g) = \Pr(Z \in A(Y), Y = y | \theta, g)$$
Marginal set likelihood

Suppose we have a set valued function $A() : \mathcal{Y} \rightarrow \sigma(\mathcal{Z})$ such that

$$g^{-1}(y) \subseteq A(y) \ \forall y, g,$$

or equivalently,

$$z \in A(g(z)) \ \forall z, g,$$

Then $\Pr(Z \in A(Y) | \theta, g) = 1$, so

$$\Pr(Y = y | \theta, g) = \Pr(Z \in A(Y), Y = y | \theta, g) = \Pr(Z \in A(y), Y = y | \theta, g)$$
Suppose we have a set valued function $A() : \mathcal{Y} \rightarrow \sigma(\mathcal{Z})$ such that

$$g^{-1}(y) \subset A(y) \, \forall y, g,$$

or equivalently,

$$z \in A(g(z)) \, \forall z, g,$$

Then $Pr(Z \in A(Y)|\theta, g) = 1$, so

$$Pr(Y = y|\theta, g) = Pr(Z \in A(Y), Y = y|\theta, g)$$

$$= Pr(Z \in A(y), Y = y|\theta, g)$$

$$= Pr(Z \in A(y)|\theta, g) \times Pr(Y = y|Z \in A(y), \theta, g)$$
Marginal set likelihood

Suppose we have a set valued function $A() : \mathcal{Y} \rightarrow \sigma(\mathcal{Z})$ such that

$$g^{-1}(y) \subseteq A(y) \quad \forall y, g,$$

or equivalently,

$$z \in A(g(z)) \quad \forall z, g,$$

Then $\Pr(Z \in A(Y)|\theta, g) = 1$, so

$$\Pr(Y = y|\theta, g) = \Pr(Z \in A(Y), Y = y|\theta, g)$$

$$= \Pr(Z \in A(y), Y = y|\theta, g)$$

$$= \Pr(Z \in A(y)|\theta, g) \times \Pr(Y = y|Z \in A(y), \theta, g)$$

$$= \Pr(Z \in A(y)|\theta) \times \Pr(Y = y|Z \in A(y), \theta, g)$$

Idea: estimate $\theta$ using only the marginal likelihood $\Pr(Z \in A(y)|\theta)$
Coarsened likelihoods

Suppose $A$ is some random set that depends on $Z$ and $g$ such that

$$\Pr(Z \in A | \theta, g) = 1 \ \forall \ \theta, g$$

Then $A$ is a *coarsening* of $Z$.

Most informative sets

Which set-valued function is most informative?

Consider the class of functions

$$\mathcal{A} = \{A() : \mathcal{Y} \to \sigma(\mathcal{Z}) \, , \, z \in A(g(z)) \, \forall z, g\}.$$ 

A marginal set likelihood could be based on any element of $\mathcal{A}$. Intuitively, we want to use the “smallest” such function $\tilde{A}()$.

**Lemma:** For each $y$, let $\tilde{A}(y) = \cap_{A} A(y)$. Then

- $\tilde{A} \in \mathcal{A}$
- $\tilde{A}(y) = \{z : y = g(z) \text{ for some } g\}$.

**Lemma:** For the copula model, $\tilde{A}(y) = \{z : z_{i1,j} < z_{i2,j} \text{ if } y_{i1,j} < y_{i2,j}\}$. 
Which set-valued function is most informative?

Consider the class of functions

\[ \mathcal{A} = \{ A() : \mathcal{Y} \to \sigma(\mathcal{Z}) , \; z \in A(g(z)) \; \forall z, g \}. \]

A marginal set likelihood could be based on any element of \( \mathcal{A} \). Intuitively, we want to use the “smallest” such function \( \tilde{A}(\cdot) \).

**Lemma:** For each \( y \), let \( \tilde{A}(y) = \bigcap \mathcal{A} A(y) \). Then

- \( \tilde{A} \in \mathcal{A} \)
- \( \tilde{A}(y) = \{ z : y = g(z) \text{ for some } g \} \).

**Lemma:** For the copula model, \( \tilde{A}(y) = \{ z : z_{i_1,j} < z_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j} \} \).
Most informative sets

Which set-valued function is most informative?

Consider the class of functions

$$\mathcal{A} = \{A() : \mathcal{Y} \to \sigma(\mathcal{Z}) , \ z \in A(g(z)) \ \forall z, g\}.$$ 

A marginal set likelihood could be based on any element of $\mathcal{A}$. Intuitively, we want to use the “smallest” such function $\tilde{A}()$.

**Lemma:** For each $y$, let $\tilde{A}(y) = \bigcap_{\mathcal{A}} A(y)$. Then

- $\tilde{A} \in \mathcal{A}$
- $\tilde{A}(y) = \{z : y = g(z) \text{ for some } g\}$.

**Lemma:** For the copula model, $\tilde{A}(y) = \{z : z_{i_1,j} < z_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j}\}$. 
Summary of Part 3

Marginal likelihoods for parameters of interest
- use part of the information in the data
- no need to estimate/construct priors for nuisance parameters

Semiparametric Bayes copula estimation
- Informative prior for parameter of interest
- Consistency for $\Sigma$ regardless of $g$
- LAN likelihood with very little information loss

But what if a marginal likelihood is not available?
Summary of Part 3

Marginal likelihoods for parameters of interest

- use part of the information in the data
- no need to estimate/construct priors for nuisance parameters

Semiparametric Bayes copula estimation

- Informative prior for parameter of interest
- Consistency for $\Sigma$ regardless of $g$
- LAN likelihood with very little information loss

But what if a marginal likelihood is not available?
Misspecified models

\[ y_1, \ldots, y_n \sim \text{iid } p_0 \]

For any density \( f \), by the SLLN,

\[
\frac{1}{n} \sum_{i=1}^{n} \log f(y_i) \rightarrow \int \log f(y) p_0(y) dy \text{ as } n \rightarrow \infty.
\]

Now consider a model \( \{ f_\theta : \theta \in \Theta \} \),

\[
\arg \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(y_i) = \hat{\theta} \rightarrow \theta_0 = \arg \max_{\theta} \int \log f_\theta(y) p_0(y) dy
\]

\[
= \arg \min_{\theta} \int \log \frac{p_0(y)}{f_\theta(y)} p_0(y) dy.
\]

\[
\sqrt{n}(\hat{\theta} - \theta_0) \sim \text{multivariate normal}(0, \Sigma_{\theta_0})
\]

(White 1982)
Misspecified models

\[ y_1, \ldots, y_n \sim \text{iid } p_0 \]

For any density \( f \), by the SLLN,

\[ \frac{1}{n} \sum_{i=1}^{n} \log f(y_i) \to \int \log f(y) p_0(y) dy \text{ as } n \to \infty. \]

Now consider a model \( \{f_\theta : \theta \in \Theta\} \),

\[
\begin{align*}
\arg \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(y_i) &= \hat{\theta} \to \theta_0 = \arg \max_{\theta} \int \log f_\theta(y) p_0(y) dy \\
&= \arg \min_{\theta} \int \log \frac{p_0(y)}{f_\theta(y)} p_0(y) dy.
\end{align*}
\]

\( \sqrt{n}(\hat{\theta} - \theta_0) \sim \text{multivariate normal}(0, \Sigma_{\theta_0}) \)

(White 1982)
Misspecified models

\[ y_1, \ldots, y_n \sim \text{iid } p_0 \]

For any density \( f \), by the SLLN,

\[
\frac{1}{n} \sum_{i=1}^{n} \log f(y_i) \to \int \log f(y) p_0(y) dy \text{ as } n \to \infty.
\]

Now consider a model \( \{f_\theta : \theta \in \Theta\} \),

\[
\arg \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(y_i) = \hat{\theta} \to \theta_0 \quad = \quad \arg \max_{\theta} \int \log f_\theta(y) p_0(y) dy
\]

\[
\quad = \quad \arg \min_{\theta} \int \log \frac{p_0(y)}{f_\theta(y)} p_0(y) dy.
\]

\[
\sqrt{n}(\hat{\theta} - \theta_0) \sim \text{multivariate normal}(0, \Sigma_{\theta_0})
\]

(White 1982)
Parameters of interest

If \( p_0 = f_{\theta_0} \) for some \( \theta_0 \in \Theta \),
1. the model is correctly specified
2. \( f_{\hat{\theta}} \to f_{\theta_0} = p_0 \)
3. \( \hat{\theta} \to \theta_0 \)

If \( p_0 \neq f_{\theta_0} \) for any \( \theta_0 \in \Theta \),
1. the model is misspecified
2. \( f_{\hat{\theta}} \to f_{\theta_0} \neq p_0 \)
3. \( \hat{\theta} \to \theta_0 \)

Can we choose a model class so that \( \theta_0 \) is something of interest?

Suppose we are interested in estimating \( \int g(y)p_0(y) \, dy \)
- we don’t need \( p_0 = f_{\theta_0} \)
- we just need \( \int g(y)p_0(y) \, dy = \int g(y)f_{\theta_0}(y) \, dy \)
Parameters of interest

If $p_0 = f_{\theta_0}$ for some $\theta_0 \in \Theta$,
1. the model is correctly specified
2. $f_\hat{\theta} \to f_{\theta_0} = p_0$
3. $\hat{\theta} \to \theta_0$

If $p_0 \neq f_{\theta_0}$ for any $\theta_0 \in \Theta$,
1. the model is misspecified
2. $f_\hat{\theta} \to f_{\theta_0} \neq p_0$
3. $\hat{\theta} \to \theta_0$

Can we choose a model class so that $\theta_0$ is something of interest?

Suppose we are interested in estimating $\int g(y)p_0(y) \, dy$

- we don’t need $p_0 = f_{\theta_0}$
- we just need $\int g(y)p_0(y) \, dy = \int g(y)f_{\theta_0}(y) \, dy$
Parameters of interest

If $p_0 = f_{\theta_0}$ for some $\theta_0 \in \Theta$,

1. the model is correctly specified
2. $f_{\hat{\theta}} \to f_{\theta_0} = p_0$
3. $\hat{\theta} \to \theta_0$

If $p_0 \neq f_{\theta_0}$ for any $\theta_0 \in \Theta$,

1. the model is misspecified
2. $f_{\hat{\theta}} \to f_{\theta_0} \neq p_0$
3. $\hat{\theta} \to \theta_0$

Can we choose a model class so that $\theta_0$ is something of interest?

Suppose we are interested in estimating $\int g(y)p_0(y) \, dy$

- we don’t need $p_0 = f_{\theta_0}$
- we just need $\int g(y)p_0(y) \, dy = \int g(y)f_{\theta_0}(y) \, dy$
Exponential families

\[ f_\theta(y) = \exp\{\theta \cdot g(y) - c(\theta)\} \]

What does \( \hat{\theta} \) converge to in this case?

\[ \theta_0 = \arg \max_\theta \int \log f_\theta(y) p_0(y) \, dy \]
\[ = \arg \max_\theta \int [\theta \cdot g(y)] p_0(y) \, dy - c(\theta) \]

Take derivatives wrt \( \theta \) and set to zero:

\[ \int g(y) p_0(y) \, dy = \frac{d}{d\theta} c(\theta)|_{\theta=\theta_0} = \int g(y) f_{\theta_0}(y) \, dy \]
Exponential families

\[ f_{\theta}(y) = \exp\{\theta \cdot g(y) - c(\theta)\} \]

What does \( \hat{\theta} \) converge to in this case?

\[ \theta_0 = \arg \max_{\theta} \int [\log f_{\theta}(y)] p_0(y) \, dy \]
\[ = \arg \max_{\theta} \int [\theta \cdot g(y)] p_0(y) \, dy - c(\theta) \]

Take derivatives wrt \( \theta \) and set to zero:

\[ \int g(y)p_0(y) \, dy = \frac{d}{d\theta} c(\theta)|_{\theta=\theta_0} = \int g(y)f_{\theta_0}(y) \, dy \]
Exponential families

\[ \lambda(\theta) = \int g(y) f_\theta(y) \, dy \quad \lambda_0 = \int g(y) p_0(y) \, dy \]

Suppose

- \( p_0 \) is true, but
- we fit \( \{ f_\theta(y) = e^{\theta \cdot g(y) - c(\theta)} : \theta \in \Theta \} \)

\[ \hat{\lambda}_n = \lambda(\hat{\theta}_n) \rightarrow \lambda(\theta_0) = \lambda(p_0) \]
\[ \sqrt{n}(\hat{\lambda} - \lambda_0) \sim \text{multivariate normal}(0, \text{Cov}_0[g(y)]) \]

Moral of the story:

\[ f_\theta(y) = \exp\{\theta_1^T \cdot g(y) + g(y)^T \theta_2 g(y) - c(\theta_1, \theta_2)\} \]

provides

- consistent estimation of \( \lambda_0 = E_{p_0}[g(y)] \)
- consistent estimation of \( \lambda_0^2 = E_{p_0}[g(y)g(y)^T] \)
- asymptotically correct confidence intervals for \( \lambda_0 \)
Nonparametric interpretation

“Nonparametric:” consistent estimation for \textit{some} functionals, regardless of $p_0$.

“Bayesian nonparametric:” Uncertainty about $p_0$?
Nonparametric interpretation from coding theory

We want to send a message \( y_1, \ldots, y_n \in \mathcal{Y} \)

- Let the relative frequencies be given by \( p_0(y) \).
- Let \( c : \mathcal{Y} \rightarrow \bigcup_{K>0} \{0, 1\}^K \) be a (prefix-free) binary code
  \[
  c(y(1)) = 0 \quad c(y(2)) = 10 \quad c(y(3)) = 110 \ldots
  \]

**Goal:** minimize the total code length

\[
\min_c \sum_{i=1}^n \text{length}[c(y_i)]
\]

**Result:** The minimal code length is given by the Shannon-Fano code

\[
\min_c \sum_{i=1}^n \text{length}[c(y_i)] = - \sum_{i=1}^n \log p_0(y_i)
\]
Nonparametric interpretation from coding theory

\[ y \in \{y(1), \ldots, y(K)\} = \mathcal{Y} \]

We want to send a message \( y_1, \ldots, y_n \in \mathcal{Y}^n \)

- Let the relative frequencies be given by \( p_0(y) \).
- Let \( c : \mathcal{Y} \to \cup_{K>0} \{0, 1\}^K \) be a (prefix-free) binary code
  \[
  c(y(1)) = 0 \quad c(y(2)) = 10 \quad c(y(3)) = 110 \ldots
  \]

Goal: minimize the total code length

\[
\min_c \sum_{i=1}^n \text{length}[c(y_i)]
\]

Result: The minimal code length is given by the Shannon-Fano code

\[
\min_c \sum_{i=1}^n \text{length}[c(y_i)] = - \sum_{i=1}^n \log p_0(y_i)
\]
**Maximum entropy and nonparametrics**

Suppose you have **no idea** what $p_0$ is. What code should you use?

**Result:**

$$\min \max \sum_{i=1}^{n} \text{length}[c(y_i)] = n \log K = -\sum_{i=1}^{n} \log (1/K)$$

The code based on the **uniform distribution** minimizes the max codelength.

What if you knew something about $p_0$, for example, $\lambda_0 = \mathbb{E}_{p_0}(g(y))$?

**Result:**

$$\min \max \sum_{i=1}^{n} \text{length}[c(y_i)] = -\sum_{i=1}^{n} \log f_{\theta_0}(y_i)$$

where $f_\theta = \exp(\theta^T g(y) - c(\theta))$ and

$$\int g(y) f_{\theta_0}(y) = \lambda_0$$
Maximum entropy and nonparametrics

Suppose you have no idea what \( p_0 \) is. What code should you use?

**Result:**

\[
\min_c \max_{y_1 \ldots y_n} \sum_{i=1}^n \text{length}[c(y_i)] = n \log K = -n \sum_{i=1}^n \log(1/K)
\]

The code based on the uniform distribution minimizes the max codelength.

What if you knew something about \( p_0 \), for example, \( \lambda_0 = E_{p_0}(g(y)) \)?

**Result:**

\[
\min_c \max_{y_1 \ldots y_n} \sum_{i=1}^n \text{length}[c(y_i)] = -\sum_{i=1}^n \log f_{\theta_0}(y_i)
\]

where \( f_\theta = \exp(\theta^T g(y) - c(\theta)) \) and

\[
\int g(y) f_{\theta_0}(y) = \lambda_0
\]
Summary of Part 4

Minimax nonparametric Bayes(?)

Parameter of interest: \( \lambda_0 = \lambda(p_0) = h(E_{p_0}[g(y)]) \)

Inference via \( \{f_\theta(y) = \exp(\theta^T g(y) - c(\theta)) \} \) provides

- a simple model focused on \( \lambda \) via \( \theta \)
  - Only prior required is on \( \theta \), or equivalently \( \lambda \)
- A “nonparametric” interpretation
  - consistent estimation of \( \lambda_0 \)
  - prior over complete density not required
  - justified via minimax criterion + lack knowledge
Summary

NPBayes methods are versatile but complicated and delicate

- opaque prior distributions/complexity penalties
- unreliable MCMC
- small sample behavior may be poor

Reality checks for new methods:

- Examine induced priors for simple functionals
- Compare results to simpler alternatives
  - robust methods that use only part of the data
  - misspecified models