Bayesian Optimal Predictive Model Selection

Ernest Fokoué

1Researcher and Consultant
QUINTESSENTIAL DIMENSIONS

Machine and Statistical Learning
Snowbird, 2006
1. Introduction to Predictive Model Selection
   - Model Specification and Optimality Criterion
   - Bayesian Predictive Optimality
   - Sparse Bayesian Learning

2. Optimal Prediction via Model Space Search
   - The Median Probability Model
   - The Prevalence Model

3. Examples, Discussion, Conclusion and Extensions
   - Examples and applications
   - Discussion, Conclusion and Extensions
1. Introduction to Predictive Model Selection
   - Model Specification and Optimality Criterion
   - Bayesian Predictive Optimality
   - Sparse Bayesian Learning

2. Optimal Prediction via Model Space Search
   - The Median Probability Model
   - The Prevalence Model

3. Examples, Discussion, Conclusion and Extensions
   - Examples and applications
   - Discussion, Conclusion and Extensions
Outline

1. Introduction to Predictive Model Selection
   - Model Specification and Optimality Criterion
     - Bayesian Predictive Optimality
     - Sparse Bayesian Learning

2. Optimal Prediction via Model Space Search
   - The Median Probability Model
   - The Prevalence Model

3. Examples, Discussion, Conclusion and Extensions
   - Examples and applications
   - Discussion, Conclusion and Extensions
INGREDIENTS FOR MODEL SPECIFICATION

- An iid data set \( \mathcal{D} \)

\[
\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, n, \ x_i \in \mathcal{X}, \ y_i \in \mathcal{Y}\}
\]

- A set of atoms

\[
\mathcal{H} = \{h_1, h_2, \ldots, h_p\}
\]

- The atomic expansion

\[
f(x_i) = \beta_0 + \sum_{j=1}^{p} \beta_j h_j(x_i)
\]

- Noisy response variable

\[
Y_i = f(x_i) + \epsilon_i
\]
DIFFERENT TYPES OF ATOMS AND EXPANSIONS

- Traditional linear model with $\mathbf{x} = (x_1, \cdots, x_p)^T \in \mathbb{R}^p$

  $$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

- Polynomial regression with $\mathbf{x} \in \mathbb{R}$

  $$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j x^j$$

- Kernel regression with $\mathbf{x} \in \mathbb{R}^p$

  $$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{n} \beta_j K(\mathbf{x}, \mathbf{x}_j)$$
The full model can be written as

\[ y = H\beta + \epsilon \]

Where the data matrix \( H \) is defined by

\[
H = \begin{bmatrix}
1 & h_1(x_1) & h_2(x_1) & \cdots & h_p(x_1) \\
1 & h_1(x_2) & h_2(x_2) & \cdots & h_p(x_2) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & h_1(x_n) & h_2(x_n) & \cdots & h_p(x_n)
\end{bmatrix}
\]

And the other elements are

\[ y = (y_1, \cdots, y_n)^T, \beta = (\beta_0, \cdots, \beta_p)^T, \epsilon = (\epsilon_1, \cdots, \epsilon_n)^T \]
Coordinate-Wise Model Description

- Define the coordinate-wise model index

\[ \mathbf{v} = (v_1, v_2, \ldots, v_p) \quad \text{where} \quad v_i \in \{0, 1\} \]

- Consider selecting from among submodels of the form

\[ M_{\mathbf{v}} : \quad \mathbf{y} = H_{\mathbf{v}} \beta_{\mathbf{v}} + \epsilon \]

Where

\[ v_i = \begin{cases} 
1 & \text{if } h_i \text{ is used by model } M_{\mathbf{v}} \\
0 & \text{otherwise} 
\end{cases} \]

- The model space \( \mathbb{M} \) with \( 2^p - 1 \) models is defined as

\[ \mathbb{M} = \{ M_{\mathbf{v}} : \quad \mathbf{v} \in \{0, 1\}^p \text{ and } \mathbf{v} \neq (0, 0, \ldots, 0) \} \]
Optimal predictive selection seeks to select from $\mathcal{M}$

$$M^*_v = \arg \min_{M_v \in \mathcal{M}} R(M_v)$$

- The risk function $R(M_v)$ here is defined by
  $$R(M_v) = \mathbb{E}[\ell(y^{\text{new}}, \hat{y}_v^{\text{new}})]$$

- The loss function is the squared error loss
  $$\ell(y^{\text{new}}, \hat{y}_v^{\text{new}}) = (y^{\text{new}} - \hat{y}_v^{\text{new}})^2$$

- The estimated prediction is given by
  $$\hat{y}_v^{\text{new}} = \sum_{j=1}^{K_v} \hat{\beta}_v^{(j)} h_v^{(j)}(x^{\text{new}}) + \hat{\beta}_0$$
1. Introduction to Predictive Model Selection
   - Model Specification and Optimality Criterion
   - Bayesian Predictive Optimality
   - Sparse Bayesian Learning

2. Optimal Prediction via Model Space Search
   - The Median Probability Model
   - The Prevalence Model

3. Examples, Discussion, Conclusion and Extensions
   - Examples and applications
   - Discussion, Conclusion and Extensions
Bayesian Model Selection and Prediction

Standard steps of Bayesian treatment

- Specify $p(M_v)$, the prior in model space
- Derive posterior probability of a model

$$p(M_v|y) \propto p(y|M_v)p(M_v)$$

- Where the marginal likelihood $p(y|M_v)$ is given by

$$p(y|M_v) = \int p(y|M_v, \theta)p(\theta|M_v)d\theta$$

Central role of $p(M_v|y)$

Model selection and prediction are based on $p(M_v|y)$
A MISLEADING INTUITION FOR MODEL SELECTION

Highest probability model

The intuition might suggest that the best predictive model is the model with the highest posterior probability

$$M_{v^*} = \arg \max_{M_v \in \mathcal{M}} p(M_v | y)$$

Some drawbacks of highest probability model

- Correct if there are only 2 models in $\mathcal{M}$.
- Requires considering all the models in $\mathcal{M}$.
- Not necessarily the best when $|\mathcal{M}| \geq 2$.
- See Babieri and Berger (2004) for details.
Optimality of BMA prediction

It is a known result that given a list of models, the Bayes Model Average (BMA) prediction

$$\hat{y}_{\text{new}} = \sum_{k=1}^{2^p - 1} p(M_{v_k} | y) \mathbb{E}[Y | M_{v_k}, x_{\text{new}}]$$

is optimal.

But ...

- Model description is lost in the averages.
- Computationally prohibitive.
When both prediction accuracy and model interpretability matter

- If accurate prediction is the only goal, then one should tolerate the computational burden and loss of model description, and adopt the BMA prediction.
- However, if there is a need to repeated predictions with the best predictor, model selection becomes the goal.

How close is the selection to BMA?

It makes sense that if we really want to select a model rather than make predictions based on BMA, the selected model should produce predictions as close to BMA predictions as possible.
Likelihood under Gaussian noise with isotropic variance $\sigma^2$

$$p(y|\beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \| y - H\beta \|^2 \right\}$$

Relevance vector prior

$$p(\beta_i|\alpha_i) = N(\beta_i; 0, \alpha_i^{-1})$$

Automatic relevance determination (ARD) hyperprior for sparsity induction

$$\alpha_i \sim \text{gamma}(a, b)$$

Marginal prior on $\beta_i$

$$p(\beta_i) = \int p(\beta_i|\alpha_i)p(\alpha_i)d\alpha_i$$
SPARSITY OF THE RVM

How does a Gaussian prior end up sparse?

- The marginal prior distribution of \( p(\beta_i) \)
  \[
p(\beta_i) = \text{Student–} t(\text{driven by } a \text{ and } b)
  \]

- The marginal prior \( p(\beta) \) is therefore a product of Student-t, and therefore a good device for sparsity.
  
  Sparsity controlled through hyperparameters \( a \) and \( b \).

- Integrating \( \beta \) out leaves us with an \( \alpha \) dependent distribution, therefore a device for controlling sparsity through ML 2.
Marginal Likelihood Under RVM

Conditional and Marginal densities

- Conditional posterior of coefficients

\[ p(\beta | \alpha, \sigma^2, y) = N(\beta; \mu, \Sigma) \]

where

\[ \Sigma = (H^T S H + A)^{-1} \]
\[ \mu = \Sigma H^T S y \]

- Where the marginal likelihood \( p(y | \alpha, \sigma^2) \) is given by

\[ p(y | \alpha, \sigma^2) = N(y; 0, S^{-1} + HA^{-1}H^T) \]
What does a true Bayesian do?

Ideally, Bayesian prediction is based on properties of the predictive distribution of the response

$$p(y^* | y) = \int p(y^* | \alpha, \sigma^2) p(\beta | y, \alpha, \sigma^2) p(\alpha, \sigma^2 | y) d\beta d\alpha d\sigma^2$$

Often though, all we hope to have is

$$p(y^* | y, \alpha, \sigma^2) = \int p(y^* | \beta, \sigma^2) p(\beta | y, \alpha, \sigma^2) d\beta$$

A serious problem ...

Very hard to find analytical expressions for $p(y^* | y)$. 
Tipping (2000) shows that ...

In our search for expression of $p(y^*|y)$, the intractability comes from

$$p(\alpha, \sigma^2 | y) = \frac{p(y|\alpha, \sigma^2)p(\alpha)p(\sigma^2)}{p(y)}$$

Instead of seeking the whole distribution $p(\alpha, \sigma^2 | y)$, we can concentrate on a crude estimate $\delta(\alpha_{MP}, \sigma^2_{MP})$ where

$$(\alpha_{MP}, \sigma^2_{MP}) = \arg \max_{\alpha, \sigma^2} p(\alpha, \sigma^2 | y)$$
Approximate Predictive Distributions

With the above posteriors at the modes

- The predictive distribution of the response is approximately

\[
p(y^* | y) \approx \int p(y^* | \beta, \sigma_{MP}^2) p(\beta | y, \alpha_{MP}, \sigma_{MP}^2) d\beta
\]

- Which is Gaussian and therefore tractable:

\[
p(\beta | y, \alpha_{MP}, \sigma_{MP}^2) = N(\beta; \mu, \Sigma)
\]

Thanks to Gaussianity, \( p(y^* | y) \approx N(y^*; \mu^*, (\sigma^2)^*) \),

\[
\mu^* = f(x_{new}; \mu)
\]

\[
(\sigma^2)^* = \sigma_{MP}^2 + h^T(x_{new}) \Sigma h(x_{new})
\]
Analysis of $f(x) = \sin(x)/x$

The sinc function is an interesting example.

- $x \in [-10, 10]$
- Sample size $n = 100$
- Noise variance $\sigma^2 = 0.01$
- Gaussian kernel used.

Figure: Sinc function
Analysis of $f(x) = \sin(x)/x$

- $x \in [-10, 10]$
- Sample size $n = 100$. $m = 1000$ repetitions.
- Noise variance $\sigma^2 = 0.1$
- Gaussian kernel used.
- Estimate noise level $= 0.093$.
- Number of relevant vectors $= 6$
- RVM regression test error (RMS): 0.0424558

Figure: RVM estimate of sinc function
What is the great appeal of RVM?

Strengths of the RVM framework
- Achieves both sparseness and fast accurate prediction.
- Works well for multivariate and univariate regression.
- Based directly on the predictive distribution.
- Flexible framework for both regression and classification.
- "Does not require a tuning parameter" ...

Drawbacks of the RVM framework
- Relevance is achieved somehow by manual pruning.
- Does not take model size into account.
- Does not provide a measure of strength of relevance.
Outline

1. Introduction to Predictive Model Selection
   - Model Specification and Optimality Criterion
   - Bayesian Predictive Optimality
   - Sparse Bayesian Learning

2. Optimal Prediction via Model Space Search
   - The Median Probability Model
   - The Prevalence Model

3. Examples, Discussion, Conclusion and Extensions
   - Examples and applications
   - Discussion, Conclusion and Extensions
**Prior Distribution Over the Model Space**

- Assuming a priori that all the atoms are equally likely, what is the prior probability that atom \( i \) is relevant, i.e., included in the model?

\[
\Pr(v_i = 1 | \pi) = \pi
\]

- Prior probability of atom \( i \)

\[
p(v_i | \pi) = \pi^{v_i} (1 - \pi)^{1 - v_i}
\]

- Assuming a priori that all the atoms are independent

\[
p(v | \pi) = \prod_{i=1}^{p} \pi^{v_i} (1 - \pi)^{1 - v_i}
\]
The model size for a model index by \( v \) is defined by

\[
K_v = \sum_{i=1}^{p} v_i
\]

Prior distribution of model size is binomial with

\[
Pr(K_v = k | \pi, p) = \binom{p}{k} \pi^k (1 - \pi)^{p-k}
\]

The prior mode of the distribution of model size

\[
k_v^\circ = \lfloor (p + 1) \pi \rfloor
\]

We can see that \( \pi \) plays a crucial role
Prior Distribution over the Model Space

- The a priori effect of $\pi$ reveals that
  
  large values of $\pi$ indicate our belief in large size models

- By the same token
  
  a small value of $\pi$ indicates our belief in small size model

- Hence, $\pi$ provides
  
  a device for controlling the grade of sparsity (parsimony)

- A key question naturally arises,
  
  How does one go about choosing $\pi$?
Noninformative Prior over Model Space

Choosing $\pi$

- Fix the value of $\pi$ for expert knowledge.
- Estimate $\pi$ using empirical Bayes on full model.
- Put a prior on $\pi$ and explore its posterior via MCMC.

Completely noninformative $p(M_v)$

By setting $\pi = 1/2$, we express our belief that our prior inclusion probability of each variable is a pure random guess. As a result

$$p(M_v) = \frac{1}{2^p}$$

which is simply a uniform prior over the $2^p$ models that constitute the model space $\mathcal{M}$. 
Let $\mathcal{M} = \{M_{v_1}, M_{v_2}, \ldots, M_{v_{2^p}}\}$ be the model space.

**Intuitively, for each atom**
- Identify all models that contain atom $h_j$.
- Compute the posterior probability of each such model.
- Compute the sum of those posterior probabilities.

**Definition**
The posterior inclusion probability for basis element $h_j$ is

$$p_j \equiv \sum_{i=1}^{2^p-1} I(v_{ij} = 1) p(M_{v_i} | y)$$
The Median Probability Model

Definition

If it exists, the median probability model, \( M_{v^*} \), is defined by

\[
v_i^* = \begin{cases} 
1 & \text{if } p_i \geq \frac{1}{2} \\
0 & \text{otherwise}
\end{cases}
\]  

(1)

Intuitively, the median probability model is made up of
- Atoms that appear more that half the time in the set of plausible models

The Median Probability model does not always exist
- The fixed "one half" cut-off may not be achieved
- Our technique uses a more flexible/adaptive cut-off.
Optimality of the Median Probability Model

Babieri and Berger (2004) show that the median probability model prediction is the best approximation to the Bayes Model Average prediction for orthogonal designs and nested designs.

Main drawbacks

- The Median Probability model does not always exist.
- The fixed "one half" cut-off may not be achieved
- Most of the search techniques used do not mix well.

Note: It is interesting to provide an extension to the MPM that has a more flexible/adaptive cut-off to guarantee existence.
Introduction to Predictive Model Selection
- Model Specification and Optimality Criterion
- Bayesian Predictive Optimality
- Sparse Bayesian Learning

Optimal Prediction via Model Space Search
- The Median Probability Model
- The Prevalence Model

Examples, Discussion, Conclusion and Extensions
- Examples and applications
- Discussion, Conclusion and Extensions
The **prevalence model**, \( M_{\nu^*} \), is defined through index \( \nu^* \), with

\[
\nu_i^{\text{prev}} = \begin{cases} 
1 & \text{if } p_i \in P^{\text{prev}} \\
0 & \text{otherwise}
\end{cases}
\]  

(2)

where

\[
P^{\text{prev}} = \{ \text{set of } k^{\text{opt}} \text{ largest values of } p_i \}
\]

and

\[
k^{\text{opt}} = \arg \max_{1 \leq k \leq p} \mathbf{p}(k|\mathbf{y})
\]  

(3)
Comparison between Median and Prevalence

- For orthogonal and nested designs, the Median probability and the Prevalence models coincide.
- For non-orthogonal designs, the Prevalence model emerges as superior and often exist where the Median fails to find any model at all.

Analytical Comparison

Elements analytical proofs in progress.
Intuitively, the prevalence model is made up of

- Atoms that appear more often in the set of plausible models of the optimal size.

The prevalence model always exists

- Intuitively better because it accounts for model size.
- Should have a connection to AIC and BIC that penalize model size.
- Later details show that the overall technique explores the support of $p(M_v|y)$ better.
**Definition**

- \( \mathcal{A} \equiv \{ a_1, a_2, \cdots, a_k \} \), where \( a_i \in \{ 1, 2, \cdots, p \} \). The elements of \( \mathcal{A} \) are called the *active* elements, because they are the ones selected by the current submodel.

- The complement of \( \mathcal{A} \) is \( \mathcal{D} = \mathcal{I} \setminus \mathcal{A} \), and contains the so-called *dormant* elements, since they are unused by the current model. \( \mathcal{D} \equiv \{ d_1, d_2, \cdots, d_{p-k} \} \), where \( d_i \in \{ 1, 2, \cdots, p \} \).

**Model uncertainty and model size uncertainty**

- **Step 1:** Birth-and-death: \((k^{(t+1)}, \mathcal{A}^{(t+1)}) \sim p(k, \mathcal{A}|\theta^{(t)}, y)\)

- **Step 2:** Gibbs sampling: \(\theta^{(t+1)} \sim p(\theta|k^{(t+1)}, \mathcal{A}^{(t+1)}, y)\)
EXTENDED PRIOR WITH UNCERTAINTY ON $k$ AND $v$

Full prior specification

- A full prior for all the unknown including model size
  \[ p(\beta, \sigma^2, \alpha, k) = p(k)p(\beta, \sigma^2, \alpha \mid k) \]
- A truncated Poisson prior on model size $k$
  \[ p(k) \propto \frac{\omega^k}{k!} e^{-\omega} \quad \text{for} \quad k = 1, \ldots, p \]
- An even full prior distribution including model indices
  \[ p(k, A, \beta, \sigma^2, \alpha) = p(k)p(A \mid k)p(\beta, \sigma^2, \alpha \mid k, A) \]

Rather than using a uniform $p(M_v)$, we instead use a locally uniform $p(A \mid k)$. Works better in the presence of collinearity.
Details of Prevalence Construction

- Initialize \( t := 0 \) and \( k^{(t)} := \lfloor p/10 \rfloor \)
- \( \mathcal{A}^{(t)} := \{ \text{sample of } k^{(t)} \text{ elements from } \{1, 2, \cdots, p\} \} \)
- Initialize inclusion probabilities \( p^{(0)} := (0, 0, \cdots, 0) \)

Repeat

- \( t := t + 1 \)
- \( \mathcal{A}^{(t)} := \text{Birth-and-Death}(\mathcal{D}^{(t-1)}, \mathcal{A}^{(t-1)}, \alpha^{(t-1)}, (\sigma^2)^{(t-1)}, y) \)
- for \( j := 1 \) to \( |\mathcal{B}| \) if \( j \in \mathcal{A}^{(t)} \) then \( p_j := p_j + 1 \) end
- \( p^{(t)} := (p_1, p_2, \cdots, p_p) \)
- \( k^{(t)} := |\mathcal{A}^{(t)}| := \text{length}(\mathcal{A}^{(t)}) \)
- \( \theta^{(t)} := \text{Gibbs-sampling}(\theta^{(t-1)}, \mathcal{A}^{(t)}, y) \)

- Until \( t = T \)
Detail of birth and death process for model search

- Initialize \( \text{time} := 0 \)
- Repeat
  - Compute \( \delta_j \) for \( j = 1, \cdots, k \); \( \delta := \sum_{j=1}^{k} \delta_j \)
  - \( \text{time} := \text{time} + \text{Exponential}(1/(\nu + \delta)) \)
  - \( \text{birth} := \text{Bernoulli} \left( \frac{\nu}{\nu + \delta} \right) \)
  - If \( \text{birth} = 1 \)
    - \( \text{in} := \text{Uniform} \left( \{ d^{(t)}_1, d^{(t)}_2, \cdots, d^{(t)}_{m-k} \} \right) \); \( k := k + 1 \)
    - \( v^{(t)}_{\text{in}} := 1; \quad A^{(t)} := A^{(t)} \cup \{ \text{in} \}; \quad D^{(t)} := D^{(t)} \setminus \{ \text{in} \} \)
  - Else
    - \( i := \text{Multinomial} (\delta_1/\delta, \cdots, \delta_k/\delta) \); \( \text{out} := a^{(t)}_i \); \( k := k - 1 \)
    - \( v^{(t)}_{\text{out}} := 0; \quad A^{(t)} := A^{(t)} \setminus \{ \text{out} \}; \quad D^{(t)} := D^{(t)} \cup \{ \text{out} \} \)
  - End;
- Until \( (\text{time} \geq \rho) \)
Birth rate and death rate of the process

Computing the death rate

- Simulate a continuous time birth and death process in discrete time using an overall constant birth rather $\nu$.
- From the local uniformity of $p(\mathcal{A} \mid k)$ and the Poisson-ness of $p(k)$, the death rate of element $i$ simplifies to

$$
\delta_i = \left[ \frac{\nu}{\omega} \right] \left[ \frac{p(y \mid k-1, \mathcal{A}_{-i}, \sigma^2, \alpha)}{p(y \mid k, \mathcal{A}, \sigma^2, \alpha)} \right]
$$

- From the normality of the likelihood function, we get

$$
\delta_i = \frac{|\sigma^2 I_n + H_v [bfA_k] H_v^T|^{\frac{1}{2}} \exp \left[ -\frac{1}{2} y^T (\sigma^2 I_n + H_v [A_{k-1}] H_v^T)^{-1} y \right]}{|\sigma^2 I_n + H_v [A_{k-1}] H_v^T|^{\frac{1}{2}} \exp \left[ -\frac{1}{2} y^T (\sigma^2 I_n + H_v [A_k] H_v^T)^{-1} y \right]}
$$
OUTLINE

1 Introduction to Predictive Model Selection
   - Model Specification and Optimality Criterion
   - Bayesian Predictive Optimality
   - Sparse Bayesian Learning

2 Optimal Prediction via Model Space Search
   - The Median Probability Model
   - The Prevalence Model

3 Examples, Discussion, Conclusion and Extensions
   - Examples and applications
   - Discussion, Conclusion and Extensions
Prevalence on the Sinc Function with Gaussian Kernel.

Figure: One of the realization of prevalence

Figure: Distribution of atom importance

Figure: Estimate distribution of $p(k|y)$
Comparison with other methods.

Summary table on the sinc function using the RBF kernel

<table>
<thead>
<tr>
<th>Prevalence vs Median</th>
<th>Prevalence</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average prediction error</td>
<td>0.227</td>
<td>0.298</td>
</tr>
<tr>
<td>Average model size</td>
<td>3.280</td>
<td>0.840</td>
</tr>
</tbody>
</table>

Table: Prevalence vs median on the sinc function

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SVR</th>
<th>RVR</th>
<th>PBR</th>
<th>SVR</th>
<th>RVR</th>
<th>PBR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinc(Gaussian)</td>
<td>0.378</td>
<td>0.326</td>
<td>0.232</td>
<td>45.2</td>
<td>6.7</td>
<td>3.5</td>
</tr>
<tr>
<td>Sinc(Uniform)</td>
<td>0.215</td>
<td>0.187</td>
<td>0.153</td>
<td>44.3</td>
<td>7.0</td>
<td>3.3</td>
</tr>
</tbody>
</table>
The following table is based on the same sinc function estimated using different orthogonal basis functions via both median and prevalence.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sine</td>
<td>2.4</td>
<td>2.4</td>
<td>0.220</td>
<td>0.220</td>
</tr>
<tr>
<td>Cosine</td>
<td>3.0</td>
<td>3.0</td>
<td>0.203</td>
<td>0.203</td>
</tr>
<tr>
<td>Legendre</td>
<td>4.2</td>
<td>4.0</td>
<td>0.197</td>
<td>0.196</td>
</tr>
<tr>
<td>Chebyshev</td>
<td>7.6</td>
<td>8.2</td>
<td>0.238</td>
<td>0.238</td>
</tr>
</tbody>
</table>

The two coincide

The clear message is that prevalence and median coincide when the design is orthogonal. Confirmation of the intuition.
PREVALENCE AND MEDIAN ON ORTHOGONAL DESIGNS.

Figure: One of the realization of prevalence

Figure: One realization of median
**Prevalence and Median on Orthogonal Designs.**

**Figure:** Distribution of atom prevalence

**Figure:** Distribution of model size
OUTLINE

1. Introduction to Predictive Model Selection
   - Model Specification and Optimality Criterion
   - Bayesian Predictive Optimality
   - Sparse Bayesian Learning

2. Optimal Prediction via Model Space Search
   - The Median Probability Model
   - The Prevalence Model

3. Examples, Discussion, Conclusion and Extensions
   - Examples and applications
   - Discussion, Conclusion and Extensions
WHAT IS THE GREAT APPEAL OF PREVALENCE ESTIMATION?

Strengths of the prevalence approach

- Achieves both sparseness and fast accurate prediction.
- Works well for multivariate and univariate regression.
- Performs a full Bayesian treatment rather than approximate.
- Provides a search technique that has very good mixing.
- Guarantee to find a solution.

Weaknesses and extensions

- The conditioning of the data matrix $H$ not always good.
- Provide a theoretical proof of computation insights.
- Consider cases with highly correlated predictor variables.