Bayesian methods in economics and finance

Linear regression: Bayesian model selection and sparsity priors
Linear regression

Model for relationship between

- (several) independent variables \( x = (x_1, \ldots, x_{D-1}) \)
- and dependent variable \( y \)

\[
y = w_0 + \sum_{i=1}^{D-1} w_i x_i + \epsilon
\]

- **Structure**: Linear relationship with parameters \( w \)
- **Noise**: Additive observation noise \( \epsilon \sim \mathcal{N}(0, \sigma_\epsilon) \)
Linear regression

Model for relationship between

- (several) independent variables \( x = (x_1, \ldots, x_{D-1}) \)
- and dependent variable \( y \)

\[
y = w_0 + \sum_{i=1}^{D-1} w_i x_i + \epsilon
\]

- Structure: Linear relationship with parameters \( w \)
- Noise: Additive observation noise \( \epsilon \sim \mathcal{N}(0, \sigma_\epsilon) \)

Convenient matrix notation

\[
y = w^T x + \epsilon
\]

where \( w, x \in \mathbb{R}^D \) and \( x_0 \equiv 1 \)
Linear regression

Example data \( \{(x_n, t_n)\}_{n=1}^{N} \)

- Collect data in matrices

\[
X = (x_1^T, \ldots, x_N^T)^T \in \mathbb{R}^{N \times D} \quad \text{and} \quad t = (t_1, \ldots, t_N)^T \in \mathbb{R}^{N \times 1}
\]

\( X \) is also called *design matrix*

- Likelihood \( p(Data|\theta) \)

\[
p(t|X, \theta) = \prod_{n=1}^{N} \mathcal{N}(t_n|w^T x_n, \sigma_\epsilon)
\]

\[
= (2\pi\sigma_\epsilon)^{-\frac{N}{2}} e^{-\frac{1}{2} \sum_{n=1}^{N} \frac{(t_n - w^T x_n)^2}{\sigma_\epsilon^2}}
\]

with parameters \( \theta = (w, \sigma_\epsilon) \)

- weights \( w = (w_0, \ldots, w_D)^T \)
- noise variance \( \sigma_\epsilon^2 \)

Note: Data are modeled as exchangeable
Ordinary least squares

Maximum likelihood solution:

$$\max_{w, \sigma} \ln p(t|X, w, \sigma) = \max_{w, \sigma} -\frac{N}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (t_n - w^T x_n)^2$$

- Find optimal weight vector $w^*$:

$$\frac{\partial}{\partial w_i} \ln p(t|X, w, \sigma) = \frac{1}{\sigma^2} \sum_{n=1}^{N} (t_n - w^T x_n) x_{in} \overset{!}{=} 0$$

$$\implies w^*^T \left( \sum_{n=1}^{N} x_n x_n^T \right) = \sum_{i=1}^{N} t_n x_n^T$$

$$w^* = (X^T X)^{-1} X^T t$$

- Minimizes the squared error $\frac{1}{2} \sum_{n=1}^{N} (t_n - w^T x_n)^2$

- $X^\dagger = (X^T X)^{-1} X^T$ is pseudo-inverse of design matrix $X$
Ordinary least squares

Maximum likelihood solution:

\[
\max_{w, \sigma, \epsilon} \ln p(t|X, w, \sigma) = \max_{w, \sigma} -\frac{N}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (t_n - w^T x_n)^2
\]

- Find optimal weight vector \(w^*\):

\[
w^* = (X^T X)^{-1} X^T t
\]

- Find optimal noise precision \(\tau_\epsilon^*\):

\[
\frac{\partial}{\partial \tau_\epsilon} \ln p(t|X, w^*, \sigma) = \frac{N}{2} \frac{1}{\tau_\epsilon} - \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T x_n)^2 \overset{!}{=} 0
\]

\[
\Rightarrow \frac{1}{\tau_\epsilon^*} = \frac{1}{N} \sum_{n=1}^{N} (t_n - w^T x_n)^2
\]
Bayesian linear regression

- Assume that noise precision $\tau_{\epsilon}$ is known
- Conjugate prior for weights $\mathbf{w}$ is Gaussian

\[ p(\mathbf{w}|0, \Sigma_0) \]

Posterior is again Gaussian with

- covariance matrix

\[ \Sigma_N = \left( \Sigma_0^{-1} + \tau_{\epsilon} \mathbf{X}^T \mathbf{X} \right)^{-1} \]

- mean

\[ \mu_N = \tau_{\epsilon} \Sigma_N \mathbf{X}^T \mathbf{t} \]
Bayesian linear regression

Posterior found by completing the square:

\[ p(w|X, t, \tau_\epsilon) \propto e^{-\frac{1}{2} \tau_\epsilon \sum_{n=1}^{N} (t_n - w^T x_n)^2} e^{-\frac{1}{2} w^T \Sigma_0^{-1} w} \]

\[ = e^{-\frac{1}{2} \tau_\epsilon (Xw - t)^T (Xw - t) + w^T \Sigma_0^{-1} w} \]

\[ \propto e^{-\frac{1}{2} w^T \left( \tau_\epsilon X^T X + \Sigma_0^{-1} \right) w + w^T \tau_\epsilon X^T t} \]

Show demo . . .
Posterior intervals

Consider the posterior $p(\theta|D)$ of a single parameter $\theta$:

- To quantify the uncertainty, define an interval $[a, b]$ which contains 95% of the posterior probability:

$$P(a < \theta < b|D) = 95\%$$

Common choices are

- Symmetric intervals, i.e. $[\hat{\theta} - x, \hat{\theta} + x]$ around the posterior mean or mode $\hat{\theta}$
- Intervals with $p(\theta = a|D) = p(\theta = b|D)$
- Intervals with $a, b$ being fixed quantiles

Also called *credible intervals, Bayesian confidence intervals*
Posterior intervals

**Example:** Estimating mean of a Gaussian $\mathcal{N}(\mu, \sigma)$ with uninformative prior:

$$P(\hat{\mu} - 1.96 \frac{\sigma}{\sqrt{N}} < \mu < \hat{\mu} - 1.96 \frac{\sigma}{\sqrt{N}}) = 95\%$$

Suggestive notation as interpretation is very different from confidence interval

- **Confidence interval:** Parameter $\mu$ is fixed and random interval *covers* $\mu$ with probability 95%
  Coverage is a repeated sampling concept
- **Credible interval:** Data, i.e. $\hat{\mu}$, are fixed and uncertain parameter lies in interval with probability 95%
Posterior intervals

▶ Frequentist interval $I(D)$ with $\alpha$ coverage:

$$P(\theta \in I(D)|\theta) = \alpha \quad \forall \theta$$

▶ With Bayesian prior $p(\theta)$ it holds:

$$\alpha = \int P(\theta \in I(D)|\theta)p(\theta)d\theta$$

$$= \int \int 1_{\theta \in I(D)}p(D, \theta)pDp\theta$$

$$= \int P(\theta \in I(D)|D)p(D)pD$$

Thus, Bayesian intervals

▶ Have good coverage on average
▶ But without worst-case guarantees, i.e. coverage for surprising data can be low
Predictive Distribution

- Predict new data point at $x_{new}$
- Predictive distribution:

$$p(t_{new} | x_{new}) = \int p(t_{new} | x_{new}, w)p(w | X, t)dw$$

Again Gaussian distribution with

- mean $\mu^T_N x_{new}$
- variance $\sigma^2_\epsilon + x^T_{new} \Sigma_N x_{new}$

*Noise variance + uncertainty about parameters*

Do not pick “best” parameters, but take uncertainty into account $\implies$ Bayesian slogan:

*Don’t optimize, integrate!*
Regularization

- Compare ML solution

\[ w^* = (X^T X)^{-1} X^T t \]

- Minimizes \( \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T x_n)^2 \)

- and posterior maximum \( (\Sigma_0^{-1} = \tau_0 I) \)

\[ \mu_N = \tau_\epsilon (\tau_0 I + \tau_\epsilon X^T X)^{-1} X^T t \]

- Minimizes regularized mean squared error

\[ \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T x_n)^2 + \frac{\lambda}{2} w^T w \]

where \( \lambda = \frac{\tau_0}{\tau_\epsilon} \)

**squared error + regularizer**
Regularization

- Compare ML solution
  
  \[ w^* = (X^T X)^{-1} X^T t \]

- Minimizes \( \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T x_n)^2 \)

- and posterior maximum \( (\Sigma_0^{-1} = \tau_0 I) \)
  
  \[ \mu_N = \tau_\epsilon (\tau_0 I + \tau_\epsilon X^T X)^{-1} X^T t \]

- Quadratic regularizer controls model complexity by shrinking weights to zero (reduces variance)

- Lasso

  - Absolute regularizer: \( \frac{\lambda}{2} \sum_{i=1}^{D} |w_i| \)
  
  - Favors sparse solutions, i.e. some weights exactly zero
  
  - Corresponds to Laplace prior: \( p(w) \propto e^{-\sum_i |w_i|} \)
Regularization

Regularized Least Squares (3)

Lasso tends to generate sparser solutions than a quadratic regularizer.
Bias-variance decomposition

Consider expected mean squared error:

\[
\mathbb{E}[(y(x; w) - t)^2] = \int \int (y(x; w) - t)^2 p(x, t) dx dt
\]

\[
= \mathbb{E}[(y(x; w) - h(x))^2] + \mathbb{E}[(h(x) - t)^2]
\]

\[
+ 2 \mathbb{E}_D[(w^T x - h(x))(h(x) - t)]
\]

\[=0 \text{ since } h(x) = \mathbb{E}[t|x]\]

- Regression function \(y(x; w) = w^T x\)
- Optimal choice is \(h(x) = \mathbb{E}[t|x]\)
- In practice \(\hat{w}\) is estimated from data set \(D\)

\[
\mathbb{E}_D[(y(x; \hat{w}) - h(x))^2] = \underbrace{(\mathbb{E}_D[y(x; \hat{w})] - h(x))^2}_{\text{bias}^2}
\]

\[
+ \underbrace{\mathbb{E}_D[(y(x; \hat{w}) - \mathbb{E}_D[y(x; \hat{w})])^2]}_{\text{variance}}
\]

\[
+ 2 \mathbb{E}_D[(y(x; \hat{w}) - \mathbb{E}_D[y(x; \hat{w})])(\mathbb{E}_D[y(x; \hat{w})] - h(x))]
\]

\[=0\]
The Bias-Variance Decomposition (5)

Example: 25 data sets from the sinusoidal, varying the degree of regularization, $\lambda$. 

![Graphs showing bias and variance decomposition](https://www.microsoft.com/en-us/research/people/cmbishop/)
The Bias-Variance Decomposition (6)

Example: 25 data sets from the sinusoidal, varying the degree of regularization, $\lambda$. 

Slide from https://www.microsoft.com/en-us/research/people/cmbishop/
Bias-variance decomposition

The Bias-Variance Decomposition (7)

Example: 25 data sets from the sinusoidal, varying the degree of regularization, $\lambda$. 

Slide from https://www.microsoft.com/en-us/research/people/cmbishop/
Bias-variance decomposition

The Bias-Variance Trade-off

From these plots, we note that an over-regularized model (large $\lambda$) will have a high bias, while an under-regularized model (small $\lambda$) will have a high variance.
Bayesian model selection

- Maximum likelihood $\max_\theta p(t|X, \theta)$ improves when number of parameters increases.
  This leads to over-fitting as model picks up noise in data.
- To achieve low expected loss we need to control
  \[ \text{bias}^2 + \text{variance} \]
  decreases with model complexity increases with model complexity

Regularization favors less flexible models by penalizing large weights

- Marginal likelihood/Evidence:
  \[ p(t|X) = \int p(t|X, w)p(w)dw \]

  - Bayesian answer to over-fitting
  - Probability of data accounting for parameter uncertainty
Bayesian model selection

Consider set of models \( \{M_i\} \), e.g. regression models with different number/subset of covariates:

- Prior probability of each model \( p(M_i) \)
- Compute posterior:

\[
p(M_i|D) \propto p(D|M_i)p(M_i)
\]

Note that each model has parameters \( w_i \) and thus

\[
p(D|M_i) = \int p(D, w_i|M_i)dw_i = \int p(D|w, M_i)p(w_i|M_i)dw_i
\]

Marginal likelihood (wrt parameters) is likelihood (wrt model)!

- Note: Marginal likelihood requires proper prior
Bayesian model selection

Two models $M_1, M_2$ are then compared based on posterior odds

$$
\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(D|M_1)}{p(D|M_2)} \cdot \frac{p(M_1)}{p(M_2)}
$$

Bayes factor · Prior odds

Interpretation of evidence from Bayes factors as suggested by Jeffreys:

<table>
<thead>
<tr>
<th>Bayes factor</th>
<th>decibans</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&lt; 1$</td>
<td>$&lt; 0$</td>
<td>negative (supports $M_2$)</td>
</tr>
<tr>
<td>$1 - 10^{1/2}$</td>
<td>0 – 5</td>
<td>barely worth mentioning</td>
</tr>
<tr>
<td>$10^{1/2} - 10$</td>
<td>5 – 10</td>
<td>substantial</td>
</tr>
<tr>
<td>$10 - 10^{3/2}$</td>
<td>10 – 15</td>
<td>strong</td>
</tr>
<tr>
<td>$10^{3/2} - 100$</td>
<td>15 – 20</td>
<td>very strong</td>
</tr>
<tr>
<td>$&gt; 100$</td>
<td>$&gt; 20$</td>
<td>decisive</td>
</tr>
</tbody>
</table>
Bayesian model selection

Marginal likelihood can be computed analytically:

- **Prior:** $p(w) = \mathcal{N}(w|0, \tau_0 I)$
- **Likelihood:** $p(t|X, w) = \mathcal{N}(t|Xw, \tau_{e} I)$
- **Evidence:**

$$
\ln p(t|X) = \ln \frac{p(t|X, w^*)p(w)^*}{p(w^*|X, t)} \quad \text{for any } w^*
$$

$$
= \frac{D}{2} \ln \tau_0 + \frac{N}{2} \ln \tau_{e} - \frac{N}{2} \ln 2\pi
$$

$$
\underbrace{w^* = \mu_N}_{w^* = \mu_N}
$$

$$
- \frac{\tau_{e}}{2} \left(X\mu_N - t\right)^T \left(X\mu_N - t\right) - \frac{\tau_0}{2} \mu_N^T \mu_N - \frac{1}{2} \ln |A|
$$

where

$$
\mu_N = \tau_{e} A^{-1} X^T t
$$

$$
A = \tau_0 I + \tau_{e} X^T X = \Sigma_N^{-1}
$$
Bayesian model selection

$0^{th}$ Order Polynomial

\[ M = 0 \]
Bayesian model selection

1st Order Polynomial

Model selection

Slide from https://www.microsoft.com/en-us/research/people/cmbishop/
Bayesian model selection

3rd Order Polynomial

Model selection

Slide from https://www.microsoft.com/en-us/research/people/cmbishop/
Bayesian model selection

9th Order Polynomial

![Graph showing a 9th Order Polynomial with points marked at x=0, x=0.5, x=1 and corresponding t values.]
Bayesian model selection

The Evidence Approximation (3)

Example: sinusoidal data, $M$\textsuperscript{th} degree polynomial,

\[ \alpha = 5 \times 10^{-3} \]

\[ \ln p(t|\alpha, \beta) \]

\[ M \]

![Graph showing the logarithm of the evidence as a function of M.](https://www.microsoft.com/en-us/research/people/cmbishop/)

Slide from https://www.microsoft.com/en-us/research/people/cmbishop/
Zellner’s g-Prior

Empirical prior with particularly easy to compute Bayes factors:

- Jeffreys prior \( p(\sigma^2_\epsilon) \propto \frac{1}{\sigma^2_\epsilon} \)
- Empirical g-prior for regression weights:

  \[
  w \sim \mathcal{N}(w_0, g\sigma^2_\epsilon(X^TX)^{-1})
  \]

- Bayes factor between model using subset \( \gamma \) of regressors and null model (intercept only):

  \[
  \frac{p(D|M_\gamma)}{p(D|M_0)} = \left(1 + g\right)^{\frac{n-|\gamma|-1}{2}} [1 + g(1 - R^2_\gamma)]^{-\frac{n-1}{2}}
  \]

where \( R^2_\gamma \) is the coefficient of determination

Note: \[
\frac{p(D|M_\gamma)}{p(D|M_{\gamma}')} = \frac{\frac{p(D|M_\gamma)}{p(D|M_0)}}{\frac{p(D|M_{\gamma}')}{p(D|M_0)}}
\]
Zellner’s g-prior

How to choose $g$?

▶ Bartlett paradoxon: General problem of uninformative prior compared to point hypothesis

$$\frac{p(\mathcal{D}|\mathcal{M}_\gamma)}{p(\mathcal{D}|\mathcal{M}_0)} \xrightarrow{g \to \infty} 0$$

▶ Information paradoxon: Consider overwhelming support for $\mathcal{M}_\gamma$, e.g. $R^2_\gamma \to 1$.

Bayes factor stays bounded as

$$\frac{p(\mathcal{D}|\mathcal{M}_\gamma)}{p(\mathcal{D}|\mathcal{M}_0)} \to (1 + g)^{\frac{n-|\gamma|-1}{2}}$$

Need to increase $g$ with number $N$ of data points:

▶ Empirical Bayes choosing $g_{ML II} = \arg\max_g p(\mathcal{D}|\mathcal{M}_\gamma)$

▶ Hyperprior for $g$:

▶ Zellner-Slow prior: $g \sim \text{Inv-Gamma}(\frac{1}{2}, \frac{N}{2})$

▶ Hyper g-prior: $\frac{g}{1+g} \sim \text{Beta}(1, \frac{\alpha}{2} - 1)$

▶ All choices are asymptotically consistent for model selection and prediction
Evidence approximation

How to choose hyperparameters, e.g. $\tau_0$?

- Bayesian ideal: Assume prior $p(\tau_0)$ and integrate
  - Often analytically intractable
  - Introduces new hyperparameters ... have to stop at some point

- Evidence approximation:
  - Optimize marginal likelihood $p(D|\tau_0)$, i.e. \textit{ML II}
  - Tends to work well in practice with little over-fitting
  - Model selection from $\{M_{\tau_0}\}_{\tau_0>0}$
Sparsity priors

Alternative to model selection:

- Fit large model with many parameters, **but** use prior that favors sparse models

  E.g. many regression weights are very small:

  - Hierarchical specification, i.e.
    
    \[
    \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \begin{pmatrix} \sigma_0^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma_D^2 \end{pmatrix}) \text{ and } \frac{1}{\sigma_i^2} \sim \text{Gamma}(\alpha, \beta)
    \]

  Evidence approximation for hyperparameters \(\sigma_1, \ldots, \sigma_D\) is called **Automatic Relevance Determination (ARD)**

- Laplace prior

  \[p(w_i) \propto e^{-|w_i|}\]

  **Bayesian LASSO**

- **Spike and Slap** prior:

  \[w_i \sim \begin{cases} 
  0 & \text{with prob. } q \\
  \mathcal{N}(0, \sigma_w) & \text{with prob. } 1 - q
  \end{cases}\]

  \[q \sim \text{Uniform}[0, 1]\]
Sparsity priors

Modern sparsity prior: Horseshoe prior

\[ w_i \sim \mathcal{N}(0, \sigma_i \sigma) \]

\[ \sigma_i \sim C^+(0, 1) \]

with the standard half-Cauchy distribution \( C^+(0, 1) \)

Figure from: C. M. Carvalho et al., *Handling Sparsity via the Horseshoe*, Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, 2009.
Sparsity priors

Modern sparsity prior: Horseshoe prior

\[ w_i \sim N(0, \sigma_i \sigma) \]
\[ \sigma_i \sim C^+(0, 1) \]

Compared to LASSO:
- Small weights: Stronger shrinkage
- Large weights: Smaller bias

Figure from: C. M. Carvalho et al., Handling Sparsity via the Horseshoe, Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, 2009.
Cross-validation

- Idea: Select model based on generalization error, i.e. best predictions on novel data
- *Cross-validation* estimates generalization error:
  - Partition data set $D$ into $K$ parts $D_1, \ldots, D_K$
  - For each $i = 1, \ldots, K$
    1. Train model on $D \setminus D_i$
    2. Evaluate model on $D_i$, e.g. using predictive distribution

  Predictive performance is estimated by average prediction error across $D_1, \ldots, D_K$

- Common choices in practice
  - $K = 10$: Good compromise between bias (due to small training set) and variance (due to small test sets)
  - $K = N$: Leaving-one-out cross-validation LOOCV
    Often efficient if data point can easily be “removed” from trained model
## Model selection summary:

<table>
<thead>
<tr>
<th>Philosophy</th>
<th>Evidence</th>
<th>Cross-validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bayesian</td>
<td>Frequentist</td>
</tr>
<tr>
<td></td>
<td>Prior ( p(\theta) ) matters</td>
<td>Prior insensitive</td>
</tr>
<tr>
<td></td>
<td>Model all data ( p(D) )</td>
<td>Predict part of data ( p(D_i</td>
</tr>
<tr>
<td></td>
<td>Often computationally demanding</td>
<td></td>
</tr>
<tr>
<td>Evaluation</td>
<td>Probability</td>
<td>Different error functions</td>
</tr>
<tr>
<td>Asymptotics</td>
<td>Consistent</td>
<td>Often inconsistent, e.g. LOOCV</td>
</tr>
<tr>
<td>Interpretation</td>
<td>Data compression</td>
<td>Prediction</td>
</tr>
</tbody>
</table>

\[
\text{BIC} = -2 \log p(D|\theta_{ML}) + D \log N
\]

\[
\text{AIC} = -2 \log p(D|\theta_{ML}) + 2D
\]