STATS271/371: Applied Bayesian Statistics

Bayesian Linear Regression

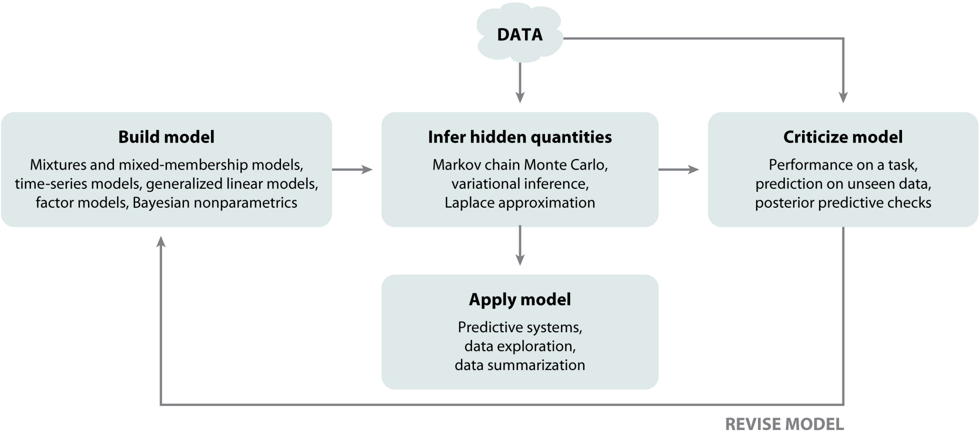
Scott Linderman

March 31, 2021

Announcements

- Please take this short survey: `https://forms.gle/urNME6zwgt1e78Rv6`
- Lecture slides are available on Canvas.
- Homework 1 will be posted on Friday (Apr 2) and due next Friday (Apr 9).

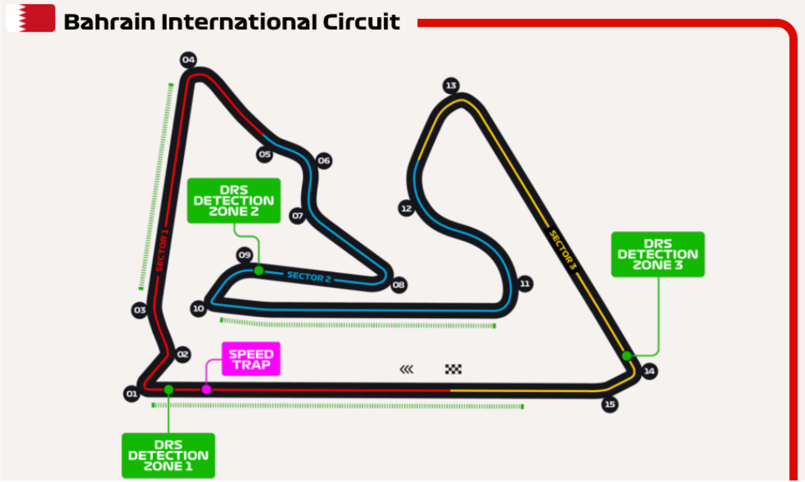
Lap 1 of Box's Loop



Blei DM. 2014.
Annu. Rev. Stat. Appl. 1:203–32

Blei, *Ann. Rev. Stat. App.* 2014.

Lap 1 of Box's Loop



https://www.formula1.com/en/racing/2021/Bahrain/Circuit.html

Bayesian Linear Regression

Our first lap around Box's loop will introduce:

- **Model:** Bayesian linear regression
- **Algorithm:** Exact posterior inference with conjugate priors
- **Criticism:** Bayesian model comparison

Notation

Let

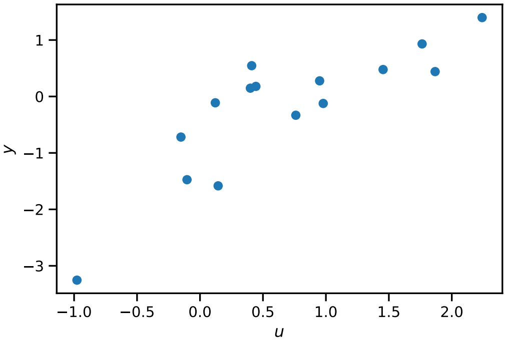
- $y_n \in \mathbb{R}$ denote the n-th *observation*
- $\boldsymbol{x}_n \in \mathbb{R}^P$ denote a the *covariates* (aka features) correspond the n-th datapoint
- $\boldsymbol{w} \in \mathbb{R}^P$ denote the *weights* of the model
- $\sigma^2 \in \mathbb{R}_+$ denote the variance of the observations

Example: Polynomial Regression

- For example, consider approximating a 1D function $y(u) : \mathbb{R} \rightarrow \mathbb{R}$ given noisy observations $\{y_n, u_n\}_{n=1}^N$.
- A priori, we don't know if the function is constant, linear, quadratic, cubic, etc.
- To fit a polynomial regression model of degree $P-1$, we can encode the inputs u_n with feature vectors,

$$\boldsymbol{x}_n = (u_n^0, u_n^1, \ldots, u_n^{P-1}) \in \mathbb{R}^P \tag{1}$$

and perform a linear regression.



Likelihood

We assume a standard Gaussian likelihood with independent noise for each datapoint,

$$p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{w}, \sigma^2) = \prod_{n=1}^N \mathcal{N}(y_n \mid \boldsymbol{w}^\top \boldsymbol{x}_n, \sigma^2) \tag{2}$$

$$= \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2\right\} \tag{3}$$

$$= \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{y_n^2}{\sigma^2} + \frac{y_n \boldsymbol{x}_n^\top \boldsymbol{w}}{\sigma^2} - \frac{1}{2}\frac{\boldsymbol{w}^\top \boldsymbol{x}_n \boldsymbol{x}_n^\top \boldsymbol{w}}{\sigma^2}\right\} \tag{4}$$

$$\propto (\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2}\left\langle \sum_{n=1}^N y_n^2, \frac{1}{\sigma^2}\right\rangle + \left\langle \sum_{n=1}^N y_n \boldsymbol{x}_n, \frac{\boldsymbol{w}}{\sigma^2}\right\rangle - \frac{1}{2}\left\langle \sum_{n=1}^N \boldsymbol{x}_n \boldsymbol{x}_n^\top, \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\sigma^2}\right\rangle\right\} \tag{5}$$

The *sufficient statistics* of the data are $(\sum_{n=1}^N y_n^2, \sum_{n=1}^N y_n \boldsymbol{x}_n, \sum_{n=1}^N \boldsymbol{x}_n \boldsymbol{x}_n^\top)$.

Aside: inner products between two matrices

The following equalities hold for the scalar quadratic form above,

$$\boldsymbol{w}^\top \boldsymbol{x}_n \boldsymbol{x}_n^\top \boldsymbol{w} = \mathrm{Tr}(\boldsymbol{w}^\top \boldsymbol{x}_n \boldsymbol{x}_n^\top \boldsymbol{w}) \tag{6}$$

$$= \mathrm{Tr}(\boldsymbol{x}_n \boldsymbol{x}_n^\top \boldsymbol{w}\boldsymbol{w}^\top) \tag{7}$$

$$= \sum_{i=1}^{P} \sum_{j=1}^{P} [\boldsymbol{x}_n \boldsymbol{x}_n^\top]_{ij} [\boldsymbol{w}\boldsymbol{w}^\top]_{ji} \tag{8}$$

$$\triangleq \langle \boldsymbol{x}_n \boldsymbol{x}_n^\top, \boldsymbol{w}\boldsymbol{w}^\top \rangle. \tag{9}$$

The inner product between two matrices $\boldsymbol{x}_n \boldsymbol{x}_n^\top$ and $\boldsymbol{w}\boldsymbol{w}^\top$ is defined by the last expression. As the sum of the element-wise product, it naturally generalizes the inner product between two vectors.

Review of maximum likelihood estimation

Before considering a Bayesian treatment, let's recall the standard maximum likelihood estimate of the parameters.

The log likelihood is,

$$
\begin{aligned}
\mathscr{L}(\boldsymbol{w}, \sigma^2) &= \log p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{w}, \sigma^2) && (10) \\
&= -\frac{N}{2} \log \sigma^2 - \frac{1}{2}\left\langle \sum_{n=1}^N y_n^2, \frac{1}{\sigma^2} \right\rangle + \left\langle \sum_{n=1}^N y_n \boldsymbol{x}_n, \frac{\boldsymbol{w}}{\sigma^2} \right\rangle - \frac{1}{2}\left\langle \sum_{n=1}^N \boldsymbol{x}_n \boldsymbol{x}_n^\top, \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\sigma^2} \right\rangle && (11)
\end{aligned}
$$

Review of maximum likelihood estimation

Taking the gradient and setting it to zero,

$$\nabla_{\boldsymbol{w}}\mathscr{L}(\boldsymbol{w},\sigma^2) = \frac{1}{\sigma^2}\sum_{n=1}^{N} y_n \boldsymbol{x}_n - \left(\frac{1}{\sigma^2}\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top\right)\boldsymbol{w} = 0 \tag{12}$$

$$\implies \boldsymbol{w}_{\text{MLE}} = \left(\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top\right)^{-1}\left(\sum_{n=1}^{N} y_n\boldsymbol{x}_n\right). \tag{13}$$

Letting

$$\boldsymbol{X} = \begin{bmatrix} - & \boldsymbol{x}_1^\top & - \\ & \vdots & \\ - & \boldsymbol{x}_N^\top & - \end{bmatrix}, \quad \text{and} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \tag{14}$$

we can write this in the more familiar form of the ordinary least squares solution,

$$\boldsymbol{w}_{\text{MLE}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}. \tag{15}$$

Review of maximum likelihood estimation II

Now let $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w}_{\text{MLE}} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ denote the predicted observations under the optimal weights. Substituting this in, we have

$$\mathscr{L}(\boldsymbol{w}_{\text{MLE}}, \sigma^2) = -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}-\hat{\boldsymbol{y}})^\top(\boldsymbol{y}-\hat{\boldsymbol{y}}) \tag{16}$$

Taking derivatives wrt $1/\sigma^2$ and setting to zero,

$$\begin{aligned}
\frac{\partial}{\partial\sigma^{-2}}\mathscr{L}(\boldsymbol{w}_{\text{MLE}}, \sigma^2) &= \frac{N}{2}\sigma^2 - \frac{1}{2}(\boldsymbol{y}-\hat{\boldsymbol{y}})^\top(\boldsymbol{y}-\hat{\boldsymbol{y}}) = 0 &(17)\\
\implies \sigma^2_{\text{MLE}} &= \frac{1}{N}(\boldsymbol{y}-\hat{\boldsymbol{y}})^\top(\boldsymbol{y}-\hat{\boldsymbol{y}}) &(18)\\
&= \frac{1}{N}(\boldsymbol{y}^\top\boldsymbol{y} - \boldsymbol{y}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}) &(19)\\
&= \frac{1}{N}\boldsymbol{y}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}, &(20)
\end{aligned}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is the *hat matrix*, which projects onto the span of the columns of $\boldsymbol{X}$.

Prior

Now consider a Bayesian treatment in which we introduce a prior on the parameters $\boldsymbol{w}$ and σ^2. Some desiderata when choosing a prior:

- It should capture intuition about the weights, like the general scale or sparsity.
- In the case where we have little prior information, it should be a broad and relatively uninformative distribution.
- All else equal, we'd prefer if it permits tractable posterior calculations.

Prior II

Let's assume we don't know much about the weights *a priori*. We'll choose a prior of the following form,

$$p(\boldsymbol{w}, \sigma^2) = \mathrm{Inv}{-}\chi^2(\sigma^2 \mid \nu, \tau^2)\,\mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{\Lambda}^{-1}), \tag{21}$$

where

- $\nu, \tau^2 \in \mathbb{R}_+$ are the degrees-of-freedom and scaling parameter, respectively, of the *inverse chi-squared distribution*.
- $\boldsymbol{\mu} \in \mathbb{R}^P$ and $\boldsymbol{\Lambda} \in \mathbb{R}^{P \times P}_{\succ 0}$ are the mean and (positive definite) precision matrix, respectively, of a *multivariate normal distribution*.

Aside: Inverse Chi-Squared Distribution

[From Wikipedia] Let s^2 be the sample mean of the squares of ν independent normal random variables with mean 0 and precision τ^2. Then $\sigma^2 = 1/s^2$ is distributed as $\mathrm{Inv}\text{—}\chi^2(\nu, \tau^2)$ and has pdf,

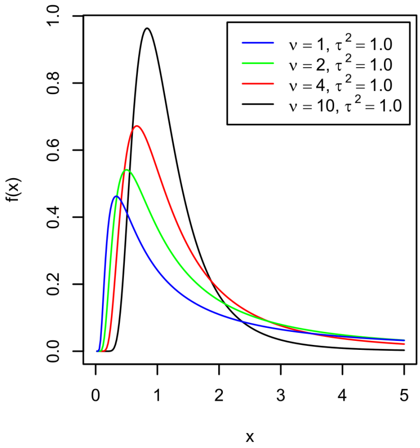
$$\mathrm{Inv}\text{—}\chi^2(\sigma^2 \mid \nu, \tau^2) = \frac{\left(\frac{\tau^2 \nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\sigma^2)^{-(1+\frac{\nu}{2})} \exp\left\{-\frac{1}{2}\left\langle \nu\tau^2, \frac{1}{\sigma^2}\right\rangle\right\} \tag{22}$$

The scaled inverse chi-squared distribution is a reparametrization of the inverse gamma distribution. Specifically, if

$$\sigma^2 \sim \mathrm{Inv}\text{—}\chi^2(\nu, \tau^2) \quad \iff \quad \sigma^2 \sim \mathrm{IGa}\left(\frac{\nu}{2}, \frac{\nu\tau^2}{2}\right). \tag{23}$$

This reparameterization is sometimes easier to work with as a conjugate prior for the variance of a Gaussian distribution.

Aside Inverse Chi-Squared Distribution II



https://en.wikipedia.org/wiki/Scaled_inverse_chi-squared_distribution

Prior density

Now expanding the prior density,

$$p(\boldsymbol{w}, \sigma^2) = \frac{\left(\frac{\tau^2 \nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} (\sigma^2)^{-(1+\frac{\nu}{2})} e^{-\frac{\nu\tau^2}{2\sigma^2}} \times (2\pi)^{-\frac{P}{2}} |\sigma^2 \boldsymbol{\Lambda}^{-1}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{w}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\boldsymbol{w}-\boldsymbol{\mu})\right\} \tag{24}$$

$$= \frac{1}{Z(\nu, \tau^2, \boldsymbol{\Lambda})} (\sigma^2)^{-(1+\frac{\nu}{2}+\frac{P}{2})} \exp\left\{-\frac{1}{2}\left\langle \nu\tau^2 + \boldsymbol{\mu}^\top \boldsymbol{\Lambda}\boldsymbol{\mu}, \frac{1}{\sigma^2}\right\rangle + \left\langle \boldsymbol{\Lambda}\boldsymbol{\mu}, \frac{\boldsymbol{w}}{\sigma^2}\right\rangle - \frac{1}{2}\left\langle \boldsymbol{\Lambda}, \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\sigma^2}\right\rangle\right\} \tag{25}$$

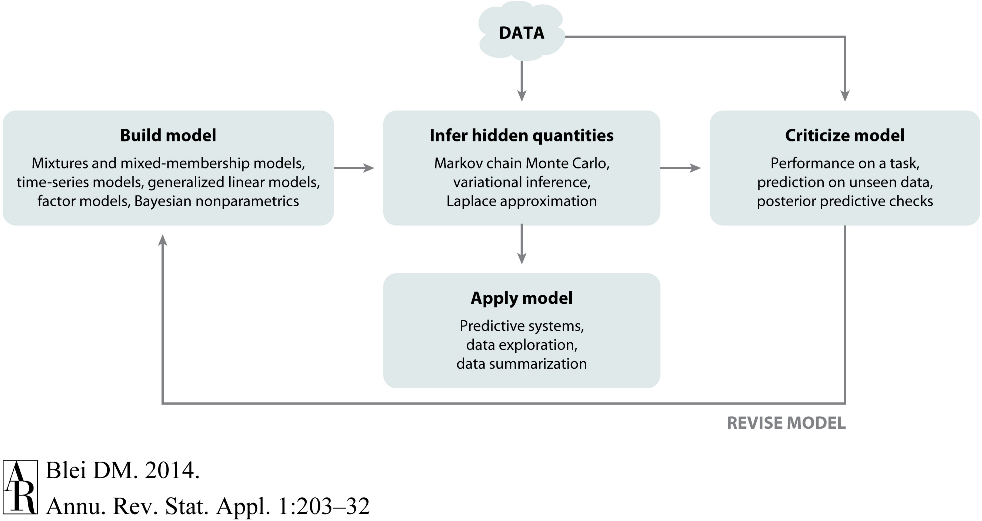
where the normalizing constant is

$$Z(\nu, \tau^2, \boldsymbol{\Lambda}) = \frac{\Gamma(\frac{\nu}{2})}{\left(\frac{\tau^2 \nu}{2}\right)^{\frac{\nu}{2}}} (2\pi)^{\frac{P}{2}} |\boldsymbol{\Lambda}|^{-\frac{1}{2}} \tag{26}$$

Properties of the prior

- **Conjugacy:** Note that the functional form is the same as in the likelihood, (5).
  - In the exponent, both are linear functions of $\frac{1}{\sigma^2}$, $\frac{\boldsymbol{w}}{\sigma^2}$, and $\frac{\boldsymbol{w}\boldsymbol{w}^\top}{\sigma^2}$.
  - This is the defining property of a *conjugate prior*, and it will lead to a closed form posterior distribution.
- **Uninformativeness:** As $\nu, \Lambda \to 0$, the prior reduces to an *improper* prior of the form $p(\boldsymbol{w}, \sigma^2) \propto (\sigma^2)^{-(1+\frac{P}{2})}$.
  - That is, it is effectively uniform in the weights and shrinking polynomially in the variance.

Box’s Loop: Infer hidden quantities



Blei, *Ann. Rev. Stat. App.* 2014.

Algorithm

- Thanks to the conjugacy of the prior, we can perform *exact* posterior inference in this model.
- Note that this is a very special case!
- The remainder of the models we'll encounter in this course will not be so nice, and we'll have to make some approximations to the posterior distribution.

Posterior Distribution

For this well behaved model we have,

$$
\begin{aligned}
p(\boldsymbol{w}, \sigma^2 \mid \{\boldsymbol{x}_n, y_n\}_{n=1}^N) &\propto p(\boldsymbol{w}, \sigma^2)\, p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{w}, \sigma^2) \\
&\propto (\sigma^2)^{-(1+\frac{\nu}{2}+\frac{P}{2}+\frac{N}{2})} \times \exp\Bigg\{ -\frac{1}{2}\Big\langle \nu\tau^2 + \boldsymbol{\mu}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} + \sum_{n=1}^N y_n^2, \frac{1}{\sigma^2} \Big\rangle \\
&\qquad + \Big\langle \boldsymbol{\Lambda}\boldsymbol{\mu} + \sum_{n=1}^N y_n \boldsymbol{x}_n, \frac{\boldsymbol{w}}{\sigma^2} \Big\rangle - \frac{1}{2}\Big\langle \boldsymbol{\Lambda} + \sum_{n=1}^N \boldsymbol{x}_n \boldsymbol{x}_n^\top, \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\sigma^2} \Big\rangle \Bigg\}
\end{aligned} \tag{27}
$$

Posterior Distribution II

Again, we see this is the same family as the prior,

$$p(\boldsymbol{w}, \sigma^2 \mid \{\boldsymbol{x}_n, y_n\}_{n=1}^N) = \text{Inv}-\chi^2(\sigma^2 \mid \nu', \tau'^2)\, \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}', \sigma^2 \boldsymbol{\Lambda}'^{-1}), \tag{28}$$

where

$$\boldsymbol{\Lambda}' = \boldsymbol{\Lambda} + \sum_{n=1}^N \boldsymbol{x}_n \boldsymbol{x}_n^\top \tag{29}$$

$$\nu' = \nu + N \tag{30}$$

$$\boldsymbol{\mu}' = \boldsymbol{\Lambda}'^{-1} \left(\boldsymbol{\Lambda}\boldsymbol{\mu} + \sum_{n=1}^N y_n \boldsymbol{x}_n \right) \tag{31}$$

$$\tau'^2 = \frac{1}{\nu'} \left(\nu\tau^2 + \boldsymbol{\mu}^\top \boldsymbol{\Lambda} \boldsymbol{\mu} + \sum_{n=1}^N y_n^2 - \boldsymbol{\mu}'^\top \boldsymbol{\Lambda}' \boldsymbol{\mu}' \right) \tag{32}$$

Uninformative limit

Consider the uninformative limit in which $\nu, \Lambda \to 0$. Then,

$$\Lambda' \to \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top \tag{33}$$

$$\nu' \to N \tag{34}$$

$$\mu' \to \left(\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top \right)^{-1} \left(\sum_{n=1}^{N} y_n \boldsymbol{x}_n \right) \tag{35}$$

$$\tau'^2 \to \frac{1}{N} \left(\sum_{n=1}^{N} y_n^2 - \left(\sum_{n=1}^{N} y_n \boldsymbol{x}_n \right)^\top \left(\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top \right)^{-1} \left(\sum_{n=1}^{N} y_n \boldsymbol{x}_n \right) \right) \tag{36}$$

Or in matrix notation

$$\Lambda' \to \boldsymbol{X}^\top \boldsymbol{X} \qquad \nu' \to N \tag{37}$$

$$\mu' \to \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1} \left(\boldsymbol{X}^\top \boldsymbol{y} \right) \qquad \tau'^2 \to \frac{1}{N} \boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{y} \tag{38}$$

Posterior Mode (aka MAP Estimate)

Under this uninformative prior, the posterior mode, aka the *maximum a posteriori* (MAP) estimate, is,

$$\boldsymbol{w}_{\text{MAP}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}^\top \boldsymbol{y}\right) \tag{39}$$

$$\sigma^2_{\text{MAP}} = \frac{\nu' \tau'^2}{\nu' + 2} = \frac{\boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{y}}{N + 2}. \tag{40}$$

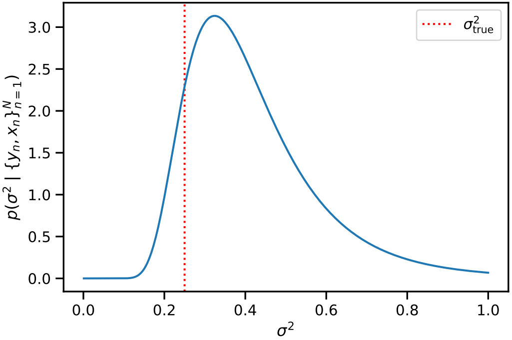
In other words, $\boldsymbol{w}_{\text{MAP}} = \boldsymbol{w}_{\text{MLE}}$ and $\sigma^2_{\text{MAP}} = \frac{N}{N+2} \sigma^2_{\text{MLE}}$.

The weights are unchanged and the variance is slightly smaller.

Posterior distribution for synthetic example

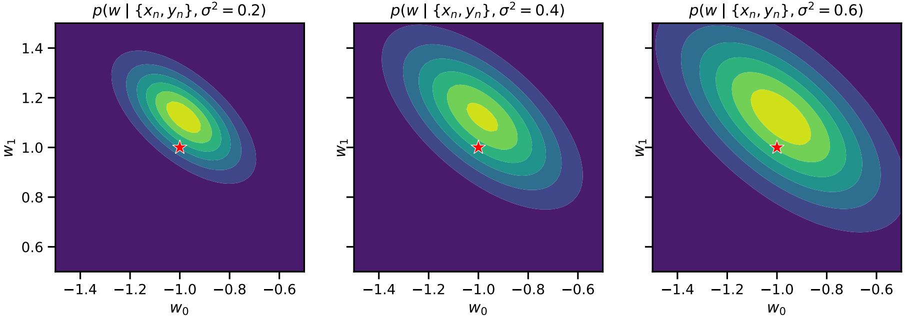
First we plot the posterior distribution of the variance,

$$p(\sigma^2 \mid \{y_n, \boldsymbol{x}_n\}_{n=1}^N) = \mathrm{IGa}(\nu', \tau'^2) \qquad (41)$$



Posterior distribution for synthetic example

Then plot $p(\boldsymbol{w} \mid \{y_n, \boldsymbol{x}_n\}_{n=1}^N, \sigma^2) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}', \sigma^2 \boldsymbol{\Lambda}'^{-1})$ for a few values of σ^2

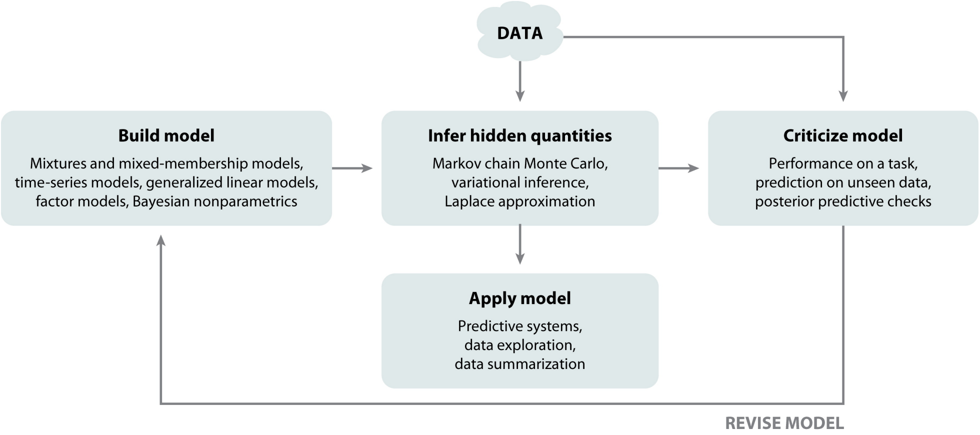


Exercise: L_2 regularization

What happens to $\boldsymbol{w}_{\text{MAP}}$ if you set $\boldsymbol{\Lambda} = \lambda \boldsymbol{I}$ for scalar $\lambda > 0$ and set $\boldsymbol{\mu} = \boldsymbol{0}$?

This is known as L_2 regularization, Tikhonov regularization, or “weight decay” in various communities.

Box’s Loop: Criticize model



Blei DM. 2014.
Annu. Rev. Stat. Appl. 1:203–32

Blei, *Ann. Rev. Stat. App.* 2014.

Model Comparison

- The marginal likelihood, aka model evidence, is a useful measure of how well a model fits the data.
- Specifically, it measures the *expected* probability assigned to the data, integrating over possible parameters under the prior,

$$
\begin{aligned}
p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N) &= \int p(\boldsymbol{w}, \sigma^2)\, p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{w}, \sigma^2)\, \mathrm{d}\boldsymbol{w}\, \mathrm{d}\sigma^2 &\quad (42)\\
&= \mathbb{E}_{p(\boldsymbol{w}, \sigma^2)}\left[p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{w}, \sigma^2)\right] &\quad (43)
\end{aligned}
$$

- If a prior distribution puts high probability on weights and variances that then assign high conditional probability to the given data, the marginal likelihood will be large.
- If the prior spreads its probability mass over a wide range of weights, it may have a lower marginal likelihood than one that concentrates mass around the weights that achieve maximal likelihood.

Marginal Likelihood

Under the conjugate prior above, we can compute the marginal likelihood in closed form,

$$
\begin{aligned}
p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N) &= \int \frac{(2\pi)^{-\frac{N}{2}}}{Z(\nu, \tau^2, \boldsymbol{\Lambda})} (\sigma^2)^{-(1+\frac{\nu'}{2}+\frac{P}{2})} \\
&\qquad \exp\left\{ -\frac{1}{2}\left\langle \nu'\tau'^2 + \boldsymbol{\mu}'^\top \boldsymbol{\Lambda}' \boldsymbol{\mu}', \frac{1}{\sigma^2} \right\rangle \right. \\
&\qquad \left. + \left\langle \boldsymbol{\Lambda}'\boldsymbol{\mu}', \frac{\boldsymbol{w}}{\sigma^2} \right\rangle - \frac{1}{2}\left\langle \boldsymbol{\Lambda}', \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\sigma^2} \right\rangle \right\} \mathrm{d}\boldsymbol{w} \, \mathrm{d}\sigma^2 && (44) \\
&= (2\pi)^{-\frac{N}{2}} \frac{Z(\nu', \tau'^2, \boldsymbol{\Lambda}')}{Z(\nu, \tau^2, \boldsymbol{\Lambda})} \int \frac{1}{Z(\nu', \tau'^2, \boldsymbol{\Lambda}')} \text{``} \qquad \ldots \qquad \text{''} \, \mathrm{d}\boldsymbol{w} \, \mathrm{d}\sigma^2 && (45) \\
&= (2\pi)^{-\frac{N}{2}} \frac{Z(\nu', \tau'^2, \boldsymbol{\Lambda}')}{Z(\nu, \tau, \boldsymbol{\Lambda})} && (46)
\end{aligned}
$$

Marginal Likelihood II

Under the conjugate prior above, we can compute the marginal likelihood in closed form,

$$p(\{y_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N) = (2\pi)^{-\frac{N}{2}} \frac{Z(\nu', \tau'^2, \boldsymbol{\Lambda}')}{Z(\nu, \tau, \boldsymbol{\Lambda})} \tag{47}$$

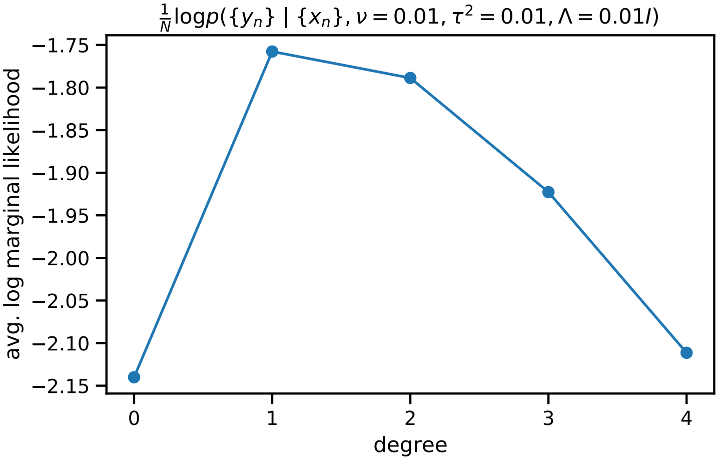
$$= (2\pi)^{-\frac{N}{2}} \frac{\Gamma(\frac{\nu'}{2})}{\Gamma(\frac{\nu}{2})} \frac{(\frac{\tau^2 \nu}{2})^{\frac{\nu}{2}}}{(\frac{\tau'^2 \nu'}{2})^{\frac{\nu'}{2}}} \frac{|\boldsymbol{\Lambda}|^{\frac{1}{2}}}{|\boldsymbol{\Lambda}'|^{\frac{1}{2}}} \tag{48}$$

Properly speaking...

Note that in order for the marginal likelihood to be meaningful, we need to have a *proper* prior distribution.

In the uninformative/improper limit, the marginal likelihood goes to zero.

Example: Using the marginal likelihood to select degree of a polynomial regression



Exercise: Unpacking the marginal likelihood

Consider selecting the degree of a polynomial regression by maximizing the marginal likelihood above. Which ratios in the marginal likelihood are growing, shrinking, or fixed, as you increase the degree P?

Preview: Posterior Predictive Distribution

- One of the main uses of regression models is to make predictions, e.g. of y_{N+1} at $\boldsymbol{x}_{N+1}$.
- In Bayesian data analysis, this is given by the *posterior predictive distribution*,

$$p(y_{N+1} \mid \boldsymbol{x}_{N+1}, \{y_n, \boldsymbol{x}_n\})_{n=1}^{N}) = \int p(y_{N+1} \mid \boldsymbol{x}_{N+1}, \boldsymbol{w}, \sigma^2)\, p(\boldsymbol{w}, \sigma^2 \mid \{y_n, \boldsymbol{x}_n\}_{n=1}^{N}\, \mathrm{d}\boldsymbol{w}\, \mathrm{d}\sigma^2 \qquad (49)$$

- Generally, we can approximate the posterior predictive distribution with Monte Carlo.
- For Bayesian linear regression with a conjugate prior, we can compute it in closed form.

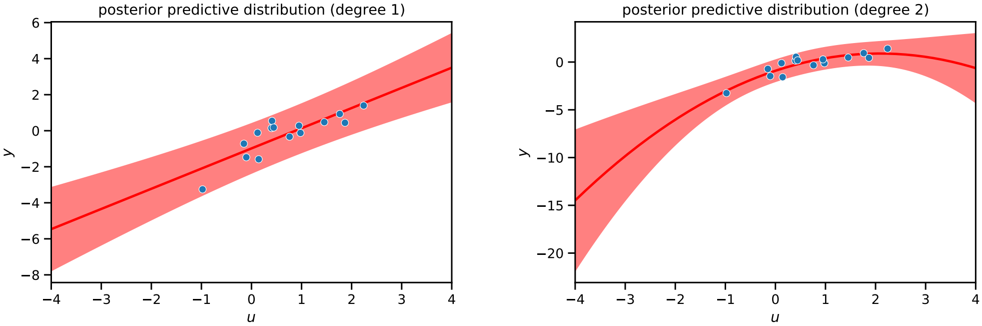
Preview: Posterior Predictive Distribution II

We have,

$$
\begin{aligned}
p(y_{N+1} \mid \boldsymbol{x}_{N+1}, \{y_n, \boldsymbol{x}_n\})_{n=1}^{N}) &= \int p(y_{N+1} \mid \boldsymbol{x}_{N+1}, \boldsymbol{w}, \sigma^2)\, p(\boldsymbol{w}, \sigma^2 \mid \{y_n, \boldsymbol{x}_n\}_{n=1}^{N}\, \mathrm{d}\boldsymbol{w}\, \mathrm{d}\sigma^2 && (50)\\
&= \int \mathcal{N}(y_{N+1} \mid \boldsymbol{w}^\top \boldsymbol{x}_{N+1}, \sigma^2)\, \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}', \sigma^2 \boldsymbol{\Lambda}'^{-1})\, \mathrm{Inv}\text{—}\chi^2(\sigma^2 \mid \nu', \tau'^2)\, \mathrm{d}\boldsymbol{w}\, \mathrm{d}\sigma^2 && (51)\\
&= \int \mathcal{N}(y_{N+1} \mid \boldsymbol{\mu}'^\top \boldsymbol{x}_{N+1}, \sigma^2(1 + \boldsymbol{x}_{N+1}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{x}_{N+1}))\, \mathrm{Inv}\text{—}\chi^2(\sigma^2 \mid \nu', \tau'^2)\, \mathrm{d}\sigma^2 && (52)\\
&= t(y_{N+1} \mid \nu', \boldsymbol{\mu}'^\top \boldsymbol{x}_{N+1}, \tau'^2(1 + \boldsymbol{x}_{N+1}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{x}_{N+1})) && (53)
\end{aligned}
$$

where $t(\cdot \mid \nu, \mu, \tau^2)$ is the density of a (generalized) *Students-t* distribution with ν degrees of freedom, location μ, and scale τ.

Preview: Posterior predictive distribution III



Bonus: Multivariate observations

Now consider multivariate observations $\boldsymbol{y}_n \in \mathbb{R}^D$ and a likelihood,

$$p(\{\boldsymbol{y}_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{W}, \boldsymbol{S}) = \prod_{n=1}^N \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{W}\boldsymbol{x}_n, \boldsymbol{S}) \tag{54}$$

where $\boldsymbol{W} \in \mathbb{R}^{D \times P}$ is now a weight *matrix* and $\boldsymbol{S} \in \mathbb{R}_{\succ 0}^{D \times D}$ is a positive definite covariance matrix.

Bonus: Multivariate observations II

Expanding the likelihood, as above, we obtain,

$$
\begin{aligned}
p(\{\boldsymbol{y}_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{W}, \boldsymbol{S}) \propto |\boldsymbol{S}|^{-\frac{N}{2}} \exp\Bigg\{ & -\frac{1}{2}\Big\langle \sum_{n=1}^N \boldsymbol{y}_n \boldsymbol{y}_n^\top, \boldsymbol{S}^{-1} \Big\rangle \\
& + \Big\langle \sum_{n=1}^N \boldsymbol{y}_n \boldsymbol{x}_n^\top, \boldsymbol{S}^{-1}\boldsymbol{W} \Big\rangle \\
& - \frac{1}{2}\Big\langle \sum_{n=1}^N \boldsymbol{x}_n \boldsymbol{x}_n^\top, \boldsymbol{W}^\top \boldsymbol{S}^{-1}\boldsymbol{W} \Big\rangle \Bigg\}
\end{aligned}
\tag{55}
$$

Conjugate prior

This is conjugate with a *matrix normal inverse Wishart* (MNIW) prior of the form,

$$
\begin{aligned}
p(\boldsymbol{W}, \boldsymbol{S}) &= \mathrm{MNIW}(\boldsymbol{W}, \boldsymbol{S} \mid \boldsymbol{\Psi}, \nu, \boldsymbol{\mu}, \boldsymbol{\Lambda}) && (56)\\
&= \mathrm{IW}(\boldsymbol{S} \mid \boldsymbol{\Psi}, \nu)\, \mathrm{MN}(\boldsymbol{W} \mid \boldsymbol{\mu}, \boldsymbol{S}, \boldsymbol{\Lambda}^{-1}). && (57)
\end{aligned}
$$

Inverse Wishart Distribution

The first the first term is an *inverse Wishart* distribution,

$$\mathrm{IW}(\boldsymbol{S} \mid \Psi, \nu) = \frac{\left(\frac{|\Psi|}{2^D}\right)^{\frac{\nu}{2}}}{\Gamma_D\left(\frac{\nu}{2}\right)} |\boldsymbol{S}|^{-(\nu+D+1)/2} \exp\left\{-\frac{1}{2}\langle\Psi, \boldsymbol{S}^{-1}\rangle\right\} \tag{58}$$

where $\Gamma_D(\cdot)$ denotes the multivariate gamma function. The inverse Wishart is a multivariate generalization of the scaled inverse chi-squared distribution.

Matrix Normal Distribution

The second term is a *matrix normal* distribution,

$$
\begin{aligned}
\mathrm{MN}(\boldsymbol{W} \mid \mu, \boldsymbol{S}, \Lambda^{-1}) &= \mathcal{N}(\mathrm{vec}(\boldsymbol{W}) \mid \mathrm{vec}(\mu), \boldsymbol{S} \otimes \Lambda^{-1}) && (59) \\
&= (2\pi)^{-\frac{DP}{2}} |\boldsymbol{S}|^{-\frac{P}{2}} |\Lambda|^{\frac{D}{2}} \exp\left\{-\tfrac{1}{2}\mathrm{Tr}\left(\Lambda(\boldsymbol{W}-\mu)^\top \boldsymbol{S}^{-1}(\boldsymbol{W}-\mu)\right)\right\} && (60) \\
&= (2\pi)^{-\frac{DP}{2}} |\boldsymbol{S}|^{-\frac{P}{2}} |\Lambda|^{\frac{D}{2}} \exp\left\{ -\frac{1}{2}\left\langle \mu\Lambda\mu^\top, \boldsymbol{S}^{-1} \right\rangle \right. \\
&\qquad\qquad + \left\langle \mu\Lambda, \boldsymbol{S}^{-1}\boldsymbol{W} \right\rangle \\
&\qquad\qquad \left. - \frac{1}{2}\left\langle \Lambda, \boldsymbol{W}^\top \boldsymbol{S}^{-1}\boldsymbol{W} \right\rangle \right\} && (61)
\end{aligned}
$$

The product of the matrix normal and inverse Wishart densities has natural parameters $\log|\boldsymbol{S}|$, $\boldsymbol{S}^{-1}$, $\boldsymbol{S}^{-1}\boldsymbol{W}$, and $\boldsymbol{W}^\top \boldsymbol{S}^{-1}\boldsymbol{W}$.

Exercise: Matrix Normal Inverse Wishart Distribution

Show that the prior used for the scalar observations is a special case of the MNIW prior.