Bayesian Approach to Machine Learning - Wireless Application

Rohit Budhiraja

Machine Learning for Wireless Communications (EE798L)

Feb. 26, 2024

Recap of last lecture and today's agenda

- Recap of last class
  - Discussed Bayesian framework for Olympic data
- Today's agenda
  - Derive Marginal likelihood for Olympic data model - Chap-4 of FCML
  - Show its application for 5G wireless systems - sparse Bayesian learning
- Extend Bayesian learning framework for non-conjugate prior and likelihood
  - Ref: Chap-4 of FCML

Bayesian treatment of Olympic data (recap)

- We treat $\mathbf{w}$ as random vector for our model $\mathbf{t} = \mathbf{X}\mathbf{w} + \epsilon$, where $\epsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_N\right)$
- From Bayes rule

$$p\left(\mathbf{w}|\mathbf{t}\right) = \frac{p\left(\mathbf{t}|\mathbf{w}\right) p\left(\mathbf{w}\right)}{p\left(\mathbf{t}\right)}$$

- Bayes rule

$$p\left(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta\right) = \frac{p\left(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2, \Delta\right) p\left(\mathbf{w}|\Delta\right)}{p\left(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta\right)} = \frac{p\left(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2\right) p\left(\mathbf{w}|\Delta\right)}{p\left(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta\right)}$$

- For our model $\mathbf{t} = \mathbf{X}\mathbf{w} + \epsilon$, likelihood is Gaussian

$$p\left(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2\right) = \mathcal{N}\left(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}_N\right)$$

- We use a Gaussian prior for $\mathbf{w}$, which conjugate to a Gaussian likelihood

$$p\left(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) = \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$$

Olympic data – Posterior calculation (recap)

- Posterior is therefore

$$p\left(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2\right) = \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}}\right)$$

  with

$$\boldsymbol{\Sigma}_{\mathbf{w}} = \left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}, \boldsymbol{\mu}_{\mathbf{w}} = \boldsymbol{\Sigma}_{\mathbf{w}}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

- If we assume prior has zero mean $\boldsymbol{\mu}_0 = \mathbf{0}$ then the posterior mean

$$\boldsymbol{\mu}_{\mathbf{w}} = \boldsymbol{\Sigma}_{\mathbf{w}}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t}\right) = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\mathbf{X}^T\mathbf{t}$$

- Posterior point estimate $\hat{\mathbf{w}} = \boldsymbol{\mu}_{\mathbf{w}}$ is the MAP estimate

Marginal likelihood for model order selection (1)

- Recall that we used cross-validation to select the order of polynomial to be used
  - Cross-validation correctly identified that dataset was generated from a third-order polynomial
- We will use marginal likelihood to determine order polynomial order for some synthetic data
- Recall the Bayes rule

$$p\left(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta\right) = \frac{p\left(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2\right) p\left(\mathbf{w}|\Delta\right)}{\int p\left(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2\right) p\left(\mathbf{w}|\Delta\right) d\mathbf{w}}$$

- Marginal likelihood for our Gaussian model is

$$p\left(\mathbf{t}|\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) = \int p\left(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2\right) p\left(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) d\mathbf{w}$$

Marginal likelihood for model order selection (2)

Theorem

Given marginal and conditional Gaussian distributions

$$p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \text{ and } p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})$$

Marginal distribution of $\mathbf{y}$ $p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x} = \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T)$

- Marginal likelihood for our Gaussian model is defined as

$$p\left(\mathbf{t}|\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) = \int p\left(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2\right) p\left(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) d\mathbf{w}$$

- With $\mathbf{x} = \mathbf{w}$ and $\mathbf{y} = \mathbf{t}$ and comparing equations below

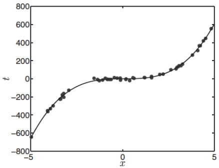
$$\begin{aligned} p\left(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) &= \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) \\ p\left(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2\right) &= \mathcal{N}\left(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}_N\right) \end{aligned}$$

- We have $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ and $\boldsymbol{\Lambda}^{-1} = \boldsymbol{\Sigma}_0$, $\mathbf{b} = \mathbf{0}$, $\mathbf{L}^{-1} = \sigma^2\mathbf{I}_N$, $\mathbf{A} = \mathbf{X}$

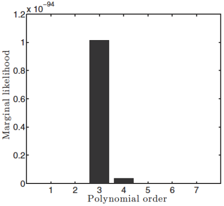
$$p\left(\mathbf{t}|\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) = \int p\left(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2\right) p\left(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) d\mathbf{w} = \mathcal{N}\left(\mathbf{X}\boldsymbol{\mu}_0, \sigma^2\mathbf{I}_N + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^T\right)$$

Marginal likelihood for model order selection (3)

- Consider a noisy third-order polynomial $t = 5x^3 - x^2 + x + \epsilon$
  - ϵ is Gaussian noise with mean zero and variance 150
- Generate data from above polynomial by uniformly picking up value from -5 to 5



(a) Noisy data from a third-order polynomial.



(b) Marginal likelihood for models of different order.

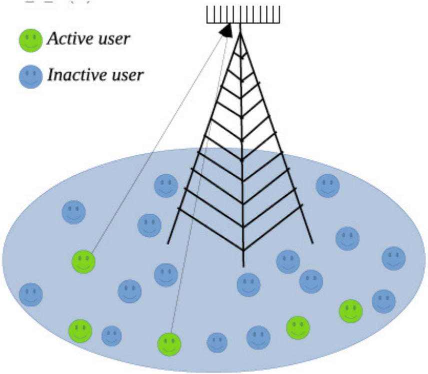
- Model the data using first to seventh-order as

$$t_n \quad = \quad w_0 + w_1 x_n + w_2 x_n^2 + \ldots + w_K x_n^K + \epsilon_n = \mathbf{w}^T \mathbf{x}_n + \epsilon_n$$

- For each model, pick a Gaussian prior with zero mean and Identity covariance matrix
  - For first-order model $\boldsymbol{\mu} = [0\ 0]^T$ and $\boldsymbol{\Sigma}_0 = \mathbf{I}_2$. For fourth model $\boldsymbol{\mu} = [0\ 0\ 0\ 0\ 0]^T$ and $\boldsymbol{\Sigma}_0 = \mathbf{I}_5$
- Evaluate marginal likelihood $p(\mathbf{t}|\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ for different priors – peaks for third order
  - Calculating marginal likelihood is very difficult and we often use cross-validation techniques

Machine learning and 5G mMTC systems (1)

- Consider a mMTC system with M single-antenna mMTC devices and N-antenna base-station (BS)



- Only few mMTC active devices transmit data which BS need to process
- BS does not know which devices are active. All active M mMTC devices transmit simultaneously
- Total number of mMTC devices $M \gg N$
- Number of active mMTC devices $K < N \ll M$

Machine learning and 5G mMTC systems (2)

- Received signal assuming all devices are active

$$
\begin{aligned}
y_1 &= h_{11}x_1 + h_{12}x_2 + \cdots + h_{1M}x_M + n_1 \\
y_2 &= h_{21}x_1 + h_{22}x_2 + \cdots + h_{2M}x_M + n_2 \\
\vdots &= \vdots \\
y_N &= h_{N1}x_1 + h_{N2}x_2 + \cdots + h_{NM}x_M + n_N \\
\mathbf{y} &= \mathbf{Hx} + \mathbf{n}
\end{aligned}
$$

- Tx signal $\mathbf{x} = [x_1, \cdots, x_M]^T$, rx signal $\mathbf{y} = [y_1, \cdots, y_N]^T$, and noise $\mathbf{n} = [n_1, \cdots, n_N]^T$
- Channel

$$
\mathbf{H} = \begin{bmatrix} h_{11} & \cdots & h_{1M} \\ \vdots & \vdots & \vdots \\ h_{N1} & \cdots & h_{NM} \end{bmatrix}
$$

Machine learning and 5G mMTC systems (3)

- Received signal assuming all devices are active

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$$

- Tx signal $\mathbf{x} = [x_1, \cdots, x_M]^T$, rx signal $\mathbf{y} = [y_1, \cdots, y_N]^T$, and noise $\mathbf{n} = [n_1, \cdots, n_N]^T$
- To recover $\mathbf{x}$ from $\mathbf{y}$, using least squares, $N \geq M$, which is not applicable here
- Recall number of active mMTC devices $K < N \ll M$
- Transmit vector $\mathbf{x}$ contains only $K \ll M$ non-zero values $\mathbf{x} = [1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0 \cdots, 0]^T$
- Transmit signal is sparse - recovery of this vector in
  - ML parlance - relevance vector machine
  - Wireless parlance - compressive sensing, sparse Bayesian learning
- We will use marginal likelihood for estimating this sparse vector $\mathbf{x}$

Sparse Bayesian learning for 5G mMTC systems (1)

- Our Olympic data model is $\mathbf{t} = \mathbf{X}\mathbf{w} + \epsilon$ for which the marginal likelihood is

$$p\left(\mathbf{t}|\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) = \mathcal{N}\left(\mathbf{X}\boldsymbol{\mu}_0, \sigma^2\mathbf{I}_N + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^T\right),$$

- Our 5G mMTC data model is $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$ for which the marginal likelihood is

$$p\left(\mathbf{y}|\mathbf{H}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) = \mathcal{N}\left(\mathbf{H}\boldsymbol{\mu}_0, \sigma^2\mathbf{I}_N + \mathbf{H}\boldsymbol{\Sigma}_0\mathbf{H}^T\right) \tag{1}$$

- We assume a Gaussian prior on $\mathbf{x}$ such that $p\left(\mathbf{x}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) = \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$ with
  - $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\boldsymbol{\Sigma}_0 = \text{diag}(\alpha_1, \cdots, \alpha_M) = \text{diag}(\boldsymbol{\alpha})$ with unknown $(\boldsymbol{\alpha})$
  - With diagonal $\boldsymbol{\Sigma}_0 = \text{diag}(\alpha_1, \cdots, \alpha_M) = \text{diag}(\boldsymbol{\alpha})$, prior is independent across entries $\boldsymbol{\alpha}$
- Such a prior as shown in next slide promotes sparsity in $\mathbf{x}$[1]
- Marginal likelihood in (1) will become

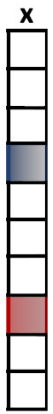
$$p\left(\mathbf{y}|\mathbf{H}, \boldsymbol{\alpha}\right) = \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_N + \mathbf{H}\text{diag}(\boldsymbol{\alpha})\mathbf{H}^T\right)$$

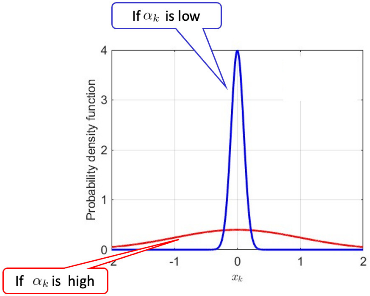
- $\boldsymbol{\alpha}$ is also called hyper-parameter, which is a parameter for parameter $\mathbf{x}$
- We assume noise variance $\sigma^2 = 1/\beta$ also to be unknown

$$p\left(\mathbf{y}|\mathbf{H}, \boldsymbol{\alpha}, \beta\right) = \mathcal{N}\left(\mathbf{0}, \beta^{-1}\mathbf{I}_N + \mathbf{H}\boldsymbol{\Sigma}_0\mathbf{H}^T\right) = \mathcal{N}(\mathbf{0}, \mathbf{C})$$

[1] Sparse Bayesian Learning and the Relevance Vector Machine, Michael E. Tipping, Journal of Machine Learning Research (2001)

How Gaussian prior promotes sparsity





- Recall we have Gaussian prior on $\mathbf{x}$ such that $p\left(\mathbf{x}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) = \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$ with
  - $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\boldsymbol{\Sigma}_0 = \text{diag}(\alpha_1, \cdots, \alpha_M)$
- If α_k is low, x_k is more likely to be close to zero
- If α_k is high, x_k is more likely to be non-zero
  - Large number of α will go to zero – posterior distribution with mean and variance zero
- With diagonal $\boldsymbol{\Sigma}_0 = \text{diag}(\alpha_1, \cdots, \alpha_M) = \text{diag}(\boldsymbol{\alpha})$, recall prior is independent across entries x_k
  - Does not capture any structured sparsity
- Similar sparsity capturing distributions - Laplace, Student-t

Sparse Bayesian learning for 5G mMTC systems (2)

- Maximize log marginal likelihood to calculate $\boldsymbol{\alpha}$ and β

$$
\begin{aligned}
p\left(\mathbf{y}|\mathbf{H},\boldsymbol{\alpha},\beta\right) &= \mathcal{N}(\mathbf{0},\mathbf{C}) = (2\pi)^{-N/2}|\mathbf{C}|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}\right\} \\
\ln p\left(\mathbf{y}|\mathbf{H},\boldsymbol{\alpha},\beta\right) &= \mathcal{L}(\boldsymbol{\alpha},\beta) = -\frac{1}{2}\left\{N\ln(2\pi)+\ln|\mathbf{C}|+\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}\right\}
\end{aligned}
$$

- Differentiate above equations wrt $\boldsymbol{\alpha}$ and β and set them to zero

$$
\begin{aligned}
\frac{\partial}{\partial\alpha_i}\ln p\left(\mathbf{y}|\mathbf{H},\boldsymbol{\alpha},\beta\right) &= 0 \\
\frac{\partial}{\partial\beta}\ln p\left(\mathbf{y}|\mathbf{H},\boldsymbol{\alpha},\beta\right) &= 0
\end{aligned}
$$

Sparse Bayesian learning for 5G+ systems (2)

- SBL is being extensively used to design 5G+ wireless systems:

- Milind Nakul, Anupama Rajoriya, and Rohit Budhiraja, "Variational Learning Algorithms For Channel Estimation in RIS-assisted mmWave Systems, IEEE Transactions on Communications", vol. 72, pp 222 - 238, Jan. 2024.
- Nishant Arya, Anupama Rajoriya, Prem Singh, and Rohit Budhiraja, "Variational Bayesian Learning Based Delay-Doppler Channel Estimator For Multi-User OTFS Systems, IEEE Communications Letters, vol. 27, pp 3355 - 3359, Dec. 2023.
- Anupama Rajoriya, and Rohit Budhiraja, "Joint AMP-SBL Algorithms For Device Activity Detection And Channel Estimation in Massive MIMO mMTC Systems, IEEE Transactions on Communications", vol. 71, pp 2136 - 2152, Apr. 2023.
- Jayanth V, Anupama Rajoriya, Nitin Gupta and Rohit Budhiraja, " Fast Correlated SBL Algorithm For Estimating Correlated Sparse Millimeter Wave Channels, IEEE Communications Letters, vol. 27, pp 1407 - 1411, May. 2023.
- Anupama Rajoriya, Alok Sharma, and Rohit Budhiraja, "Covariance-Free Variational Bayesian Learning For Correlated Block Sparse Signals, IEEE Communications Letters, vol. 27, pp 966 - 970, Mar. 2023.
- Anupama Rajoriya, Rohit Budhiraja and Lajos Hanzo, "Centralized and Decentralized Channel Estimation in FDD Multi-User Massive MIMO Systems, IEEE Transactions on Vehicular Technology, vol. 71, pp 7325 - 7342, Jul. 2022.
- Anupama Rajoriya, Syed Rukhsana and Rohit Budhiraja, "Centralized And Decentralized Active User Detection And Channel Estimation in mMTC", IEEE Transactions on Communications", vol. 70, pp 1759 - 1776, Mar. 2022.

Appendix

- Updates of $\boldsymbol{\alpha}$ and β

Marginal likelihood and sparse Bayesian learning (2)

- Recall Posterior is given as

$$p\left(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2\right) = \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{w}},\boldsymbol{\Sigma}_{\mathbf{w}}\right)$$

  with

$$\boldsymbol{\Sigma}_{\mathbf{x}} = \left(\beta\mathbf{H}^T\mathbf{H} + \mathrm{diag}(\boldsymbol{\alpha})\right)^{-1}$$

  and

$$\boldsymbol{\mu}_{\mathbf{x}} = \boldsymbol{\Sigma}_{\mathbf{x}}\left(\frac{1}{\sigma^2}\mathbf{H}^T\mathbf{y} + \mathrm{diag}(\boldsymbol{\alpha})\boldsymbol{\mu}_0\right) = \beta\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^T\mathbf{t}$$

- Posterior mean and covariance expression will be used multiple times while deriving the updates

Simplification of log marginal likelihood expression (1)

- Marginal likelihood to calculate $\boldsymbol{\alpha}$ and β is as follows

$$p\left(\mathbf{y}|\mathbf{H},\beta,\boldsymbol{\alpha}\right) = \mathcal{N}(\mathbf{0},\mathbf{C}) = (2\pi)^{-N/2}|\mathbf{C}|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}\right\}$$

- Log of marginal likelihood

$$\mathcal{L}(\boldsymbol{\alpha},\beta) = -\frac{1}{2}\left\{N\ln(2\pi) + \underbrace{\ln|\mathbf{C}|}_{T_1} + \underbrace{\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}}_{T_2}\right\} \qquad (2)$$

- Recall that $|\mathbf{C}| = |\beta^{-1}\mathbf{I} + \mathbf{H}\boldsymbol{\Sigma}_0\mathbf{H}^T|$

$$\begin{aligned}
|\beta^{-1}\mathbf{I} + \mathbf{H}\boldsymbol{\Sigma}_0\mathbf{H}^T| &= |\beta^{-1}\mathbf{I}||\mathbf{I} + \beta\mathbf{H}\boldsymbol{\Sigma}_0\mathbf{H}^T| = |\beta^{-1}\mathbf{I}||\mathbf{I} + \beta\boldsymbol{\Sigma}_0\mathbf{H}^T\mathbf{H}| = |\beta^{-1}\mathbf{I}||\boldsymbol{\Sigma}_0||\boldsymbol{\Sigma}_0^{-1} + \beta\mathbf{H}^T \\
\Rightarrow |\beta^{-1}\mathbf{I} + \mathbf{H}\boldsymbol{\Sigma}_0\mathbf{H}^T| &= |\beta^{-1}\mathbf{I}||\boldsymbol{\Sigma}_0||\boldsymbol{\Sigma}_0^{-1} + \beta\mathbf{H}^T\mathbf{H}| \\
|\boldsymbol{\Sigma}_0^{-1}||\underbrace{\beta^{-1}\mathbf{I} + \mathbf{H}\boldsymbol{\Sigma}_0\mathbf{H}^T}_{\mathbf{C}}| &= |\beta^{-1}\mathbf{I}||\underbrace{\boldsymbol{\Sigma}_0^{-1} + \beta\mathbf{H}^T\mathbf{H}}_{\boldsymbol{\Sigma}_x^{-1}}| \\
\Rightarrow |\mathbf{C}| &= \frac{|\beta^{-1}\mathbf{I}||\boldsymbol{\Sigma}_x^{-1}|}{|\boldsymbol{\Sigma}_0|} = \frac{|\beta^{-1}\mathbf{I}||\boldsymbol{\Sigma}_x^{-1}|}{|\mathrm{diag}(\boldsymbol{\alpha})|}
\end{aligned}$$

Simplification of log marginal likelihood expression (2)

- Recall $|\mathbf{C}| = \frac{|\beta^{-1}\mathbf{I}||\mathbf{\Sigma}_x^{-1}|}{|\text{diag}(\boldsymbol{\alpha})|}$. We next simplify T_1 from (2)

$$T_1 = \ln|\mathbf{C}| = -N\ln\beta - \ln|\mathbf{\Sigma}_x| - \sum_{i=1}^{N}\ln\alpha_i$$

- Woodbury identity : $(\mathbf{A}+\mathbf{UDV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{D}^{-1}+\mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}$
- With $\mathbf{A} = \beta^{-1}\mathbf{I}$, $\mathbf{U} = \mathbf{H}, \mathbf{D} = \text{diag}(\boldsymbol{\alpha})$ and $\mathbf{V} = \mathbf{H}^T$, we equivalently express $\mathbf{C}^{-1}$

$$\mathbf{C}^{-1} = \left(\beta^{-1}\mathbf{I} + \mathbf{H}(\text{diag}(\boldsymbol{\alpha}))^{-1}\mathbf{H}^T\right)^{-1} = \beta\mathbf{I} - \beta\mathbf{H}\left(\text{diag}(\boldsymbol{\alpha}) + \beta\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\beta = \beta\mathbf{I} - \beta\mathbf{H}\mathbf{\Sigma}_x\mathbf{H}^T\beta$$

- We next simplify T_2 as follows

$$\begin{aligned} T_2 &= \mathbf{y}^T\mathbf{C}^{-1}\mathbf{y} = \beta\mathbf{y}^T\mathbf{y} - \beta\mathbf{y}^T\mathbf{H}\mathbf{\Sigma}_x\mathbf{H}^T\mathbf{y}\beta = \beta\mathbf{y}^T(\mathbf{y} - \mathbf{H}\underbrace{\mathbf{\Sigma}_x\mathbf{H}^T\mathbf{y}\beta}_{\boldsymbol{\mu}_x}) = \beta\mathbf{y}^T(\mathbf{y} - \mathbf{H}\boldsymbol{\mu}_x) \\ &= \beta\mathbf{y}^T\mathbf{y} - \beta\mathbf{y}^T\mathbf{H}\boldsymbol{\mu}_x \underbrace{-\beta\mathbf{y}^T\mathbf{H}\boldsymbol{\mu}_x + \beta\boldsymbol{\mu}_x^T\mathbf{H}^T\mathbf{H}\boldsymbol{\mu}_x}_{\text{adding and subtracting for completing the squares}} + \beta\mathbf{y}^T\mathbf{H}\boldsymbol{\mu}_x - \beta\boldsymbol{\mu}_x^T\mathbf{H}^T\mathbf{H}\boldsymbol{\mu}_x \\ &= \beta||\mathbf{y}^T\mathbf{y} - \mathbf{H}\boldsymbol{\mu}_x||^2 + \beta\mathbf{y}^T\mathbf{H}\boldsymbol{\mu}_x - \beta\boldsymbol{\mu}_x^T\mathbf{H}^T\mathbf{H}\boldsymbol{\mu}_x \end{aligned}$$

Simplification of log marginal likelihood expression (3)

- We re-express $\beta \mathbf{y}^T \mathbf{H} \boldsymbol{\mu}_x$ as

$$\begin{aligned} \beta \boldsymbol{\Sigma}_x \mathbf{H}^T \mathbf{y} &= \boldsymbol{\mu}_x \\ \beta \mathbf{H}^T \mathbf{y} &= \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\mu}_x \\ \beta \boldsymbol{\mu}_x^T \mathbf{H}^T \mathbf{y} = \beta \mathbf{y}^T \mathbf{H} \boldsymbol{\mu}_x &= \boldsymbol{\mu}_x^T \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\mu}_x \end{aligned} \tag{3}$$

- Using (3), we have

$$\begin{aligned} T_2 &= \beta ||\mathbf{y} - \mathbf{H}\boldsymbol{\mu}_x||^2 + \boldsymbol{\mu}_x^T \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\mu}_x - \beta \boldsymbol{\mu}_x^T \mathbf{H}^T \mathbf{H} \boldsymbol{\mu}_x \\ &= \beta ||\mathbf{y} - \mathbf{H}\boldsymbol{\mu}_x||^2 + \boldsymbol{\mu}_x^T (\boldsymbol{\Sigma}_x^{-1} - \beta \mathbf{H}^T \mathbf{H}) \boldsymbol{\mu}_x \qquad \text{where} \quad \boldsymbol{\Sigma}_x^{-1} = \text{diag}(\boldsymbol{\alpha}) + \beta \mathbf{H}^T \mathbf{H} \\ &= \beta ||\mathbf{y} - \mathbf{H}\boldsymbol{\mu}_x||^2 + \boldsymbol{\mu}_x^T (\text{diag}(\boldsymbol{\alpha}) + \beta \mathbf{H}^T \mathbf{H} - \beta \mathbf{H}^T \mathbf{H}) \boldsymbol{\mu}_x \\ &= \beta ||\mathbf{y} - \mathbf{H}\boldsymbol{\mu}_x||^2 + \boldsymbol{\mu}_x^T \text{diag}(\boldsymbol{\alpha}) \boldsymbol{\mu}_x \\ &= \beta ||\mathbf{y} - \mathbf{H}\boldsymbol{\mu}_x||^2 + \boldsymbol{\mu}_x^T \text{diag}(\boldsymbol{\alpha}) \boldsymbol{\mu}_x \end{aligned}$$

- Using T_1 and T_2 we rewrite (2) as follows

$$\mathcal{L}(\alpha, \beta) = -\frac{1}{2} \left\{ N \ln(2\pi) - N \ln \beta - \ln |\boldsymbol{\Sigma}_x| - \sum_{i=1}^{N} \ln \alpha_i + \beta ||\mathbf{y} - \mathbf{H}\boldsymbol{\mu}_x||^2 + \boldsymbol{\mu}_x^T \text{diag}(\boldsymbol{\alpha}) \boldsymbol{\mu}_x \right\}$$

Calculation of value of α (1)

- Differentiating log likelihood with respect to α_i

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \alpha_i} &= \frac{1}{2}\frac{\partial}{\partial \alpha_i}(\ln|\boldsymbol{\Sigma}_x|) + \frac{1}{2\alpha_i} - \frac{1}{2}\frac{\partial}{\partial \alpha_i}(\boldsymbol{\mu}_x^T \text{diag}(\boldsymbol{\alpha})\boldsymbol{\mu}_x) \\ \frac{\partial \mathcal{L}}{\partial \alpha_i} &= \frac{1}{2}\underbrace{\frac{\partial}{\partial \alpha_i}(\ln|\boldsymbol{\Sigma}_x|)}_{D_1} + \frac{1}{2\alpha_i} - \frac{1}{2}(\boldsymbol{\mu}_x(i))^2\end{aligned} \tag{4}$$

- Next

$$D_1 = \frac{\partial}{\partial \alpha_i}(\ln|\boldsymbol{\Sigma}_x|) \stackrel{(a)}{=} -\frac{\partial}{\partial \alpha_i}(\ln|\text{diag}(\boldsymbol{\alpha}) + \beta \mathbf{H}^T\mathbf{H}|) \stackrel{(b)}{=} -\text{Tr}(\boldsymbol{\Sigma}_x(i,i)) = -\boldsymbol{\Sigma}_x(i,i)$$

- Equality (a) uses

$$\boldsymbol{\Sigma}_x = (\text{diag}(\boldsymbol{\alpha}) + \beta \mathbf{H}^T\mathbf{H})^{-1}$$

- Equality (b) uses the property

$$\frac{\partial}{\partial x}(\ln|\mathbf{A}|) = \text{Tr}(\mathbf{A}^{-1}\frac{\partial}{\partial x}\mathbf{A})$$

Calculation of value of α (2)

- Substituting D_1 in (4), we get

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \alpha_i} &= -\frac{1}{2}\boldsymbol{\Sigma}_x(i,i) + \frac{1}{2\alpha_i} - \frac{1}{2}(\boldsymbol{\mu}_x(i))^2 = 0 \\
\frac{1}{2\alpha_i} &= \frac{1}{2}[\boldsymbol{\Sigma}_x(i,i) + (\boldsymbol{\mu}_x(i))^2] \\
\alpha_i &= \frac{1-\alpha_i\boldsymbol{\Sigma}_x(i,i)}{(\boldsymbol{\mu}_x(i))^2} \\
\alpha_i &= \frac{1-\gamma_i}{(\boldsymbol{\mu}_x(i))^2}
\end{aligned}
$$

- where $\gamma_i = \alpha_i\boldsymbol{\Sigma}_x(i,i)$

Calculation of value of β (1)

- $$\mathcal{L}(\alpha,\beta) = -\frac{1}{2}[N\ln(2\pi) - N\ln\beta - \ln|\boldsymbol{\Sigma}_x| - \sum_{i=1}^{N}\ln\alpha_i + \beta||\mathbf{y} - \mathbf{H}\boldsymbol{\mu}_x||^2 + \boldsymbol{\mu}_x^T\mathrm{diag}(\alpha)\boldsymbol{\mu}_x]$$

- $$\frac{\partial\mathcal{L}(\alpha,\beta)}{\partial\beta} = -\frac{1}{2}\Big\{-\frac{N}{\beta} - \underbrace{\frac{\partial}{\partial\beta}\ln|\boldsymbol{\Sigma}_x|}_{D_2} + ||\mathbf{y} - \mathbf{H}\boldsymbol{\mu}_x||^2\Big\} = 0$$

- $$\begin{aligned} D_2 &= \frac{\partial}{\partial\beta}\ln|\boldsymbol{\Sigma}_x| = -\frac{\partial}{\partial\beta}\ln|\boldsymbol{\Sigma}_x^{-1}| = -\frac{\partial}{\partial\beta}\ln|\mathrm{diag}(\boldsymbol{\alpha}) + \beta\mathbf{H}^T\mathbf{H}| \\ &= -\mathrm{Tr}(\boldsymbol{\Sigma}_x\mathbf{H}^T\mathbf{H}) = -\mathrm{Tr}\left(\boldsymbol{\Sigma}_x\mathbf{H}^T\mathbf{H} + \beta^{-1}\boldsymbol{\Sigma}_x\mathrm{diag}(\alpha) - \beta^{-1}\boldsymbol{\Sigma}_x\mathrm{diag}(\alpha)\right) \\ &= -\mathrm{Tr}\Big(\boldsymbol{\Sigma}_x\underbrace{\left(\mathbf{H}^T\mathbf{H}\beta + \mathrm{diag}(\alpha)\right)}_{\boldsymbol{\Sigma}_x^{-1}}\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}_x\mathrm{diag}(\alpha)\Big) \\ &= -\mathrm{Tr}(\beta^{-1}\mathbf{I} - \beta^{-1}\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_0^{-1}) = -\frac{1}{\beta}\mathrm{Tr}\left(\mathbf{I} - \boldsymbol{\Sigma}_x\mathrm{diag}(\alpha)\right) = -\frac{1}{\beta}(N - \sum_{i=1}^{N}\gamma_i) \end{aligned}$$

Calculation of value of β (2)

- $$\begin{aligned}\frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \beta} &= -\frac{1}{2}\Big\{-\frac{N}{\beta} - \underbrace{\frac{\partial}{\partial \beta}\ln|\mathbf{\Sigma}_x|}_{D_2} + ||\mathbf{y}-\mathbf{H}\boldsymbol{\mu}_x||^2\Big\} = 0 \\ &= -\frac{1}{2}\left\{-\frac{N}{\beta} + \frac{1}{\beta}(N - \sum_{i=1}^{N}\gamma_i) + ||\mathbf{y}-\mathbf{H}\boldsymbol{\mu}_x||^2\right\} = 0 \\ \sigma^2 = \frac{1}{\beta} &= \frac{||\mathbf{y}-\mathbf{H}\boldsymbol{\mu}_x||^2}{\sum_{i=1}^{N}\gamma_i}\end{aligned}$$