Bayesian Ordinal Regression

April 28, 2022

1 Introduction

Ordinal regression is a type of regression analysis used for predicting an ordinal variable, i.e., a variable whose values exists on an arbitrary scale where only the relative ordering between different values is significant. In [1], Gaussian Process model was presented for the ordinal regression models. The following sections describes this model.

2 Latent variable model for ordinal regression

Assume we have a training set $\mathcal{D}$ of n (IID) observations, $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) | i = 1, .., n\}$, where $\boldsymbol{x} \in \mathbb{R}^d$ denotes an input vector (covariates) of dimension $\mathcal{D}$ and y denotes an ordinal response variable on a scale $0, .., K$. Let Z be a latent variable that underlies the generation of the ordinal responses

$$Z = f(\boldsymbol{x}) + \epsilon \tag{1}$$

where $f(\boldsymbol{x})$ is a zero-mean Gaussian process with covariance function $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}')$ and ϵ is zero mean Gaussian noise with variance σ_ϵ^2 . I.e.,

$$p(Z|f(\boldsymbol{x}), \sigma_\epsilon^2) = \mathcal{N}(f(\boldsymbol{x}), \sigma_\epsilon^2). \tag{2}$$

The response variable y results from an "incomplete measurements" of Z,

$$y = \begin{cases} 1, & Z \leq \eta_1, \\ 2, & \eta_1 \leq Z \leq \eta_2 \\ \vdots & \\ K & \eta_{K-1} \leq Z. \end{cases} \tag{3}$$

Defining $\eta_0 = -\infty$ and $\eta_K = \infty$, the above can be summarized as $y = k$ if (and only if) $\eta_{k-1} \leq Z \leq \eta_k$.

Denote with $\boldsymbol{\theta}$ the parameters vector including the thresholds, $\{\eta_1, ..., \eta_{K-1}\}$, the noise level σ_ϵ^2 and the covariance function parameters. The probability that y equals k

$$
\begin{aligned}
P_r(y = k|f(\boldsymbol{x}), \boldsymbol{\theta}) = P_r(\eta_{k-1} \leq Z \leq \eta_k | f(\boldsymbol{x}), \sigma_\epsilon^2) = \\
= \int_{\eta_{k-1}}^{\eta_k} p(z|f(\boldsymbol{x}), \sigma_\epsilon^2) dz = \Phi\left(\frac{\eta_k - f(\boldsymbol{x})}{\sigma_\epsilon}\right) - \Phi\left(\frac{\eta_{k-1} - f(\boldsymbol{x})}{\sigma_\epsilon}\right) \quad (4)
\end{aligned}
$$

where Φ is the cumulative distribution function of the Gaussian distribution. For $k = 1$ and $k = K$ we obtain,

$$
\begin{aligned}
P_r(y = 0|f(\boldsymbol{x}), \boldsymbol{\theta}) &= \Phi\left(\frac{\eta_k - f(\boldsymbol{x})}{\sigma_\epsilon}\right) \\
P_r(y = K|f(\boldsymbol{x}), \boldsymbol{\theta}) &= 1 - \Phi\left(\frac{\eta_{k-1} - f(\boldsymbol{x})}{\sigma_\epsilon}\right)
\end{aligned}
$$

The conditional distribution of y is now given by

$$
\begin{aligned}
p(y|f(\boldsymbol{x}), \boldsymbol{\theta}) &= \prod_{i=1}^{K} P_r(y = k|f(\boldsymbol{x}), \boldsymbol{\theta})^{[y=k]} = \\
&= \prod_{k=1}^{K} \left[\Phi\left(\frac{\eta_k - f(\boldsymbol{x})}{\sigma_\epsilon}\right) - \Phi\left(\frac{\eta_{k-1} - f(\boldsymbol{x})}{\sigma_\epsilon}\right)\right]^{[y=k]} = \\
&= \Phi\left(\frac{\eta_y - f(\boldsymbol{x})}{\sigma_\epsilon}\right) - \Phi\left(\frac{\eta_{y-1} - f(\boldsymbol{x})}{\sigma_\epsilon}\right)
\end{aligned}
$$

using the Iverson bracket $[y = k]$. Then, the likelihood of the ordinal model can now be stated as

$$
p(\mathcal{D}|\boldsymbol{f}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|f(\boldsymbol{x}_i), \boldsymbol{\theta}), \quad (5)
$$

where $\boldsymbol{f} = [f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), ..., f(\boldsymbol{x}_n)]^T$.

3 Full Bayesian treatment

In the full Bayesian treatment, we must assign priors for both to $\boldsymbol{f}$ and $\boldsymbol{\theta}$. The posterior probability can then be written as

$$
\begin{aligned}
p(\boldsymbol{f}, \boldsymbol{\theta}|\mathcal{D}) &= \frac{p(\mathcal{D}|\boldsymbol{f}, \boldsymbol{\theta})p(\boldsymbol{f}, \boldsymbol{\theta})}{p(\mathcal{D})} \underset{f, \boldsymbol{\theta}\ iid}{=} \\
&= \frac{p(\mathcal{D}|\boldsymbol{f}, \boldsymbol{\theta})p(\boldsymbol{f})p(\boldsymbol{\theta})}{p(\mathcal{D})} \quad (6)
\end{aligned}
$$

where $p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{f}, \boldsymbol{\theta})p(\boldsymbol{f})p(\boldsymbol{\theta})d\boldsymbol{f}d\boldsymbol{\theta}$. The posterior predictive distribution is given by

$$
p(y_*|\boldsymbol{x}_*, \mathcal{D}) = \int p(y_*|f(\boldsymbol{x}_*), \boldsymbol{\theta})p(\boldsymbol{f}, \boldsymbol{\theta}|\mathcal{D})d\boldsymbol{f}d\boldsymbol{\theta} \quad (7)
$$

Computing the posterior distribution is analytically intractable and most often Monte Carlo methods are used to obtain approximations.

3.1 Prior specification

The prior probability for $\boldsymbol{f}$ is a multivariate Gaussian

$$p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{0}, K) \tag{8}$$

where K is $n \times n$ covariance matrix whose ij-element equals $\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

To set priors to the thresholds, $\{\eta_1, ..., \eta_{K-1}\}$ that enforce increasing order and positivity we use the following definition

$$\eta_j = \eta_1 + \sum_{l=2}^{j} \log \Delta_l, \quad j = 2, .., K-1. \tag{9}$$

Now we can assign normal prior over the parameters $\{\eta_1, \log \Delta_2, ..., \log \Delta_{K-1}\}$.

4 Partial Bayesian treatment

In a full Bayesian treatment the parameters $\boldsymbol{\theta}$ must be integrated over the $\boldsymbol{\theta}$-space. An alternative solution is to find a point estimate for $\boldsymbol{\theta}$. This results in a Bayesian framework conditional on the parameters $\boldsymbol{\theta}$

$$p(\boldsymbol{f}|\mathcal{D}, \boldsymbol{\theta}) = \frac{p(\mathcal{D}|\boldsymbol{f}, \boldsymbol{\theta})p(\boldsymbol{f})}{p(\mathcal{D}|\boldsymbol{\theta})}. \tag{10}$$

A point estimate for $\boldsymbol{\theta}$ can be either computed by maximizing the evidence $p(\mathcal{D}|\boldsymbol{\theta})$ (ML estimator) or by maximizing the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ (MAP estimator), where $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. The evidence (which is required for the estimating $\boldsymbol{\theta}$) is given by a high dimensional integral,

$$p(\mathcal{D}|\boldsymbol{\theta}) = \int p(\mathcal{D}|\boldsymbol{f}, \boldsymbol{\theta})p(\boldsymbol{f})d\boldsymbol{f}, \tag{11}$$

which is analytically intractable. A popular approach is to approximate the posterior, $p(\boldsymbol{f}|\mathcal{D}, \boldsymbol{\theta})$ as a Gaussian (e.g. using Laplace approximation), and then the evidence can be calculates using explicit formula. The MLE or MAP estimation of $\boldsymbol{\theta}$ can then be obtained using gradient-based optimization methods.

4.1 Laplace Approximation

In this section, we develop the Laplace approximation of the posterior $p(\boldsymbol{f}|\mathcal{D},\boldsymbol{\theta})$ at the maximum a-posteriori (MAP) estimate. Denotes with $\Psi(\boldsymbol{f})$ the (unnormalized) negative log of the posterior

$$\Psi(\boldsymbol{f}) = -\log p(\mathcal{D}|\boldsymbol{f},\boldsymbol{\theta})p(\boldsymbol{f}) \tag{12}$$

The Laplace approximation refers to using the Taylor expansion of $\Psi(\boldsymbol{f})$ at the MAP point and retaining the terms up to the second order. This is equivalent to approximate the posterior with the following Gaussian distribution

$$p(\boldsymbol{f}|\mathcal{D},\boldsymbol{\theta}) = \mathcal{N}(\hat{\boldsymbol{f}}, H^{-1}) \tag{13}$$

where $\hat{\boldsymbol{f}}$ is the MAP estimate of the posterior and H is the Hessian matrix of $\Psi(\boldsymbol{f})$ evaluated at $\hat{\boldsymbol{f}}$

$$\hat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}} \Psi(\boldsymbol{f}) \tag{14}$$

$$H = \frac{\partial^2 \Psi(\boldsymbol{f})}{\partial \boldsymbol{f} \partial \boldsymbol{f}^T}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}}. \tag{15}$$

Using Eq. (5),(8) we obtain the following expression for $\Psi(\boldsymbol{f})$

$$\Psi(\boldsymbol{f}) = \sum_{i=1}^{n} \ell(y_i, f(\boldsymbol{x}_i), \boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{f}^T K^{-1} \boldsymbol{f} + \frac{n}{2}\log 2\pi + \frac{1}{2}\log|K| \tag{16}$$

where

$$\begin{aligned} \ell(y_i, f(\boldsymbol{x}_i), \boldsymbol{\theta}) &= -\log p(y_i|f(\boldsymbol{x}_i), \boldsymbol{\theta}) \\ &= -\log\left[\Phi\left(\frac{\eta_{y_i} - f(\boldsymbol{x}_i)}{\sigma_\epsilon}\right) - \Phi\left(\frac{\eta_{y_i-1} - f(\boldsymbol{x}_i)}{\sigma_\epsilon}\right)\right] \end{aligned} \tag{17}$$

In Appendix A we show that the Hessian matrix is given by

$$H = \Lambda + K^{-1}$$

where Λ is diagonal matrix whose ii-elements is $\frac{\partial^2 \ell(y_i, f_i, \boldsymbol{\theta})}{\partial f_i^2}$ given as in (31). We further show (Appendix A) that H is positive define, hence the optimization problem involve in finding the MAP estimate is convex and have a unique solution. The resulting Gaussian approximation is given by

$$p(\boldsymbol{f}|\mathcal{D},\boldsymbol{\theta}) = \mathcal{N}(\hat{\boldsymbol{f}}, (\Lambda_{\text{MAP}} + K^{-1})^{-1}), \tag{18}$$

where Λ_{MAP} denotes Λ at the MAP estimate. The Second order Taylor expansion of $\Psi(\boldsymbol{f})$ at the MAP estimate is given by

$$\hat{\Psi}(\boldsymbol{f}) = \Psi(\hat{\boldsymbol{f}}) + \frac{1}{2}(\boldsymbol{f} - \hat{\boldsymbol{f}})^T(\Lambda_{\text{MAP}} + K^{-1})(\boldsymbol{f} - \hat{\boldsymbol{f}}) \tag{19}$$

Table 1: Algorithm for model adaptation using the MAP approach with Laplace approximation

Initialization	choose a favorite gradient-descent optimization package.
	select the starting point $\boldsymbol{\theta}$ for the optimization package.
Looping	While the optimization package requests evidence/gradient evaluation at $\boldsymbol{\theta}$
	(1) Find the MAP estimate by solving the convex optimization problem (14)
	(2) Evaluate the **negative** log-evidence (20) at the MAP estimate
	(3) Calculate the gradients with respect to $\boldsymbol{\theta}$ (Appendix B).
	(4) feed the evidence and gradients to the optimization package.
Exit	Return the optimal $\boldsymbol{\theta}$ found by optimization package

4.2 Model Adaptation

To obtain a point estimate for $\boldsymbol{\theta}$ we must compute the evidence $p(\mathcal{D}|\boldsymbol{\theta})$ as defined in (11). Using (12) we obtain

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\theta}) &= \int p(\mathcal{D}|\boldsymbol{f}, \boldsymbol{\theta})p(\boldsymbol{f})d\boldsymbol{f} \\ &= \int \exp -\Psi(\boldsymbol{f}). \end{aligned}$$

Replacing $\Psi(\boldsymbol{f})$ with the approximation $\hat{\Psi}(\boldsymbol{f})$ (19) we obtain,

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\theta}) \approx \int \exp -\hat{\Psi}(\boldsymbol{f}) = \\ &= \exp\left[-\Psi(\hat{\boldsymbol{f}})\right] \int \exp\left[-\frac{1}{2}(\boldsymbol{f}-\hat{\boldsymbol{f}})^T(\Lambda_{\mathrm{MAP}} + K^{-1})(\boldsymbol{f}-\hat{\boldsymbol{f}})\right] d\boldsymbol{f} \\ &= \exp\left[-\Psi(\hat{\boldsymbol{f}})\right] (2\pi)^{n/2}|\Lambda_{\mathrm{MAP}} + K^{-1}|^{-1/2} = \\ &= \end{aligned}$$

Hence, the log evidence is given by

$$\begin{aligned} \log p(\mathcal{D}|\boldsymbol{\theta}) &\approx -\sum_{i=1}^{n} \ell(y_i, \hat{f}_i, \boldsymbol{\theta}) - \frac{1}{2}\hat{\boldsymbol{f}}^T K^{-1}\hat{\boldsymbol{f}} - \frac{1}{2}\log|K||\Lambda_{\mathrm{MAP}} + K^{-1}| + C \\ &= -\sum_{i=1}^{n} \ell(y_i, \hat{f}_i, \boldsymbol{\theta}) - \frac{1}{2}\hat{\boldsymbol{f}}^T K^{-1}\hat{\boldsymbol{f}} - \frac{1}{2}\log|I + K\Lambda_{\mathrm{MAP}}| + C \end{aligned} \tag{20}$$

The gradient of (20) with respect to the hyper-parameters $\boldsymbol{\theta}$ can be derived analytically (Appendix B). The outline of algorithm for model adaptation is described in Table 1.

4.3 Prediction

At the optimal hyper-parameters we inferred, denoted as $\hat{\boldsymbol{\theta}}$, let us take a test case $\boldsymbol{x}_*$ for which the ordinal response variable y_* is unknown. To predict y_* we first predict the latent variable $f_* \triangleq f(\boldsymbol{x}_*)$. The posterior predictive distribution $p(f_*|\boldsymbol{x}_*\mathcal{D}, \hat{\boldsymbol{\theta}})$ can be written as

$$p(f_*|\boldsymbol{x}_*, \mathcal{D}, \hat{\boldsymbol{\theta}}) = \int p(f_*|x_*, \boldsymbol{f})p(\boldsymbol{f}|\mathcal{D}, \hat{\boldsymbol{\theta}})d\boldsymbol{f}$$

where $p(\boldsymbol{f}|\mathcal{D}, \hat{\boldsymbol{\theta}})$ is the posterior approximation (18) and $p(f_*|x_*\boldsymbol{f})$ is the prior conditional distribution of f_* given $\boldsymbol{f}$. Since the posterior and the prior are Gaussian, the posterior predictive distribution is also Gaussian

$$p(f_*|\boldsymbol{x}_*\mathcal{D}, \hat{\boldsymbol{\theta}}) = \mathcal{N}(\mu_x, \sigma_x^2).$$

Furthermore, the mean μ_x and the variance σ_x^2 satisfies

$$\mu_x = \boldsymbol{k}_*^T K^{-1} \boldsymbol{\mu}_p \tag{21}$$
$$\sigma_x^2 = \mathcal{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^T (K^{-1} - K^{-1}\Sigma_p K^{-1})\boldsymbol{k}_* \tag{22}$$

where $\boldsymbol{k}_* = [\mathcal{K}(\boldsymbol{x}_1, \boldsymbol{x}_*), \mathcal{K}(\boldsymbol{x}_2, \boldsymbol{x}_*), ..., \mathcal{K}(\boldsymbol{x}_n, \boldsymbol{x}_*)]$, and $\boldsymbol{\mu}_p, \Sigma_p$ are the mean vector and covariance matrix of the posterior. Using (18) and the matrix inverse lemma we obtain

$$\begin{aligned} \mu_x &= \boldsymbol{k}_*^T K^{-1} \hat{\boldsymbol{f}}, \\ \sigma_x^2 &= \mathcal{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^T (K + \Lambda_{\text{MAP}}^{-1})^{-1} \boldsymbol{k}_*. \end{aligned}$$

The predictive distribution over ordinal targets y_x is

$$\begin{aligned} p(y_x|\boldsymbol{x}_*, \mathcal{D}, \hat{\boldsymbol{\theta}}) &= \int p(y_x|\boldsymbol{x}_*, f_*, \hat{\boldsymbol{\theta}}) p(f_*|\boldsymbol{x}_*, \mathcal{D}, \hat{\boldsymbol{\theta}}) = \\ &= \Phi\left(\frac{\eta_{y_x} - \mu_x}{\sqrt{\sigma_\epsilon^2 + \sigma_x^2}}\right) - \Phi\left(\frac{\eta_{y_x - 1} - \mu_x}{\sqrt{\sigma_\epsilon^2 + \sigma_x^2}}\right). \end{aligned} \tag{23}$$

Appendix A - Hessian Matrix

The Hessian matrix of $\Psi(\boldsymbol{f})$ (16) is given by

$$\begin{aligned} H = \frac{\partial^2 \Psi}{\partial \boldsymbol{f} \partial \boldsymbol{f}^T} &= \frac{\partial}{\partial \boldsymbol{f} \partial \boldsymbol{f}^T} \sum_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i), \boldsymbol{\theta}) + \frac{\partial^2}{\partial \boldsymbol{f} \partial \boldsymbol{f}^T} \frac{1}{2} \boldsymbol{f}^T K^{-1} \boldsymbol{f} \\ &= \frac{\partial^2}{\partial \boldsymbol{f} \partial \boldsymbol{f}^T} \sum_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i), \boldsymbol{\theta}) + K^{-1} = \Lambda + K^{-1} \end{aligned} \tag{24}$$

where Λ is the Hessian matrix of $\sum_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i), \boldsymbol{\theta})$. The ij-element of Λ is given by

$$\Lambda_{ij} = \frac{\partial^2}{\partial f(\boldsymbol{x}_i) f(\boldsymbol{x}_j)} \sum_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i), \boldsymbol{\theta}) = \begin{cases} \frac{\partial^2 \ell(y_i, f(\boldsymbol{x}_i), \boldsymbol{\theta})}{\partial f(\boldsymbol{x}_i)^2} & i = j \\ 0 & i \neq j \end{cases}. \tag{25}$$

That is, Λ is a diagonal matrix whose ii element equals $\frac{\partial^2 \ell(y_i, f(\boldsymbol{x}_i), \boldsymbol{\theta})}{\partial f(\boldsymbol{x}_i)^2}$. For better clarity we use $f_i = f(\boldsymbol{x}_i)$ hereinafter. Using Eq. (17) we obtain,

$$\begin{aligned}
\Lambda_{ii} &= \frac{\partial^2 \ell(y_i, f_i, \boldsymbol{\theta})}{\partial f_i^2} = \\
&= -\frac{\partial^2}{\partial f_i^2} \log \left[\Phi\left(\frac{\eta_{y_i} - f_i}{\sigma_\epsilon}\right) - \Phi\left(\frac{\eta_{y_i-1} - f_i}{\sigma_\epsilon}\right)\right], \\
&= -\frac{\partial^2}{\partial f_i^2} \log \left[\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)\right]
\end{aligned} \tag{26}$$

where $z_1^i = \frac{\eta_{y_i} - f_i}{\sigma_\epsilon}$ and $z_2^i = \frac{\eta_{y_i-1} - f_i}{\sigma_\epsilon}$. The first derivative of $\log\left[\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)\right]$ with respect to f_i is given by

$$\begin{aligned}
\frac{\partial}{\partial f_i} \log\left[\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)\right] &= \frac{\Phi'\left(z_1^i\right) - \Phi'\left(z_2^i\right)}{\left[\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)\right]} \\
&= -\frac{1}{\sigma_\varepsilon} \frac{\mathcal{N}(z_1^i, 0, 1) - \mathcal{N}(z_2^i, 0, 1)}{\left[\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)\right]}
\end{aligned} \tag{27}$$

The second derivative with respect to f_i is given by

$$\begin{aligned}
\frac{\partial^2}{\partial f_i^2} \log\left[\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)\right] &= -\frac{1}{\sigma_\varepsilon} \frac{\partial}{\partial f_i} \frac{\mathcal{N}(z_1^i, 0, 1) - \mathcal{N}(z_2^i, 0, 1)}{\left[\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)\right]} \\
&= -\frac{1}{\sigma_\varepsilon} \frac{\partial}{\partial f_i} \frac{h(f_i)}{k(f_i)} = -\frac{1}{\sigma_\varepsilon} \frac{h'(f_i)k(f_i) - h(f_i)k'(f_i)}{k(f_i)^2}
\end{aligned} \tag{28}$$

where

$$\begin{aligned}
h(f_i) &= \mathcal{N}(z_1^i, 0, 1) - \mathcal{N}(z_2^i, 0, 1) \\
h'(f_i) &= \frac{\partial}{\partial f_i}\left[\mathcal{N}(z_1^i, 0, 1) - \mathcal{N}(z_2^i, 0, 1)\right] \\
&= -\mathcal{N}(z_1^i, 0, 1) z_1^i \frac{\partial z_1^i}{\partial f_i} + \mathcal{N}(z_2^i, 0, 1) z_2^i \frac{\partial z_2^i}{\partial f_i} \\
&= \frac{1}{\sigma_\varepsilon}\left[z_1^i \mathcal{N}(z_1^i, 0, 1) - z_2^i \mathcal{N}(z_2^i, 0, 1)\right] \\
k(f_i) &= \Phi\left(z_1^i\right) - \Phi\left(z_2^i\right) \\
k'(f_i) &= \mathcal{N}(z_1^i, 0, 1) \frac{\partial z_1^i}{\partial f_i} - \mathcal{N}(z_2^i, 0, 1) \frac{\partial z_2^i}{\partial f_i}
\end{aligned}$$

Thus we obtain,

$$\begin{aligned}
h'(f_i)k(f_i) - h(f_i)k'(f_i) &= \frac{1}{\sigma_\varepsilon}\left[z_1^i \mathcal{N}(z_1^i, 0, 1) - z_2^i \mathcal{N}(z_2^i, 0, 1)\right]\left[\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)\right] \\
&+ \frac{1}{\sigma_\varepsilon}\left[\mathcal{N}(z_1^i, 0, 1) - \mathcal{N}(z_2^i, 0, 1)\right]^2
\end{aligned} \tag{29}$$

and hence,

$$\frac{\partial^2}{\partial f_i^2}\log\left[\Phi\left(z_1^i\right)-\Phi\left(z_2^i\right)\right] = -\frac{1}{\sigma_\varepsilon^2}\left[\frac{z_1^i\mathcal{N}(z_1^i,0,1)-z_2^i\mathcal{N}(z_2^i,0,1)}{\Phi\left(z_1^i\right)-\Phi\left(z_2^i\right)}\right] - \frac{1}{\sigma_\varepsilon^2}\left[\frac{\mathcal{N}(z_1^i,0,1)-\mathcal{N}(z_2^i,0,1)}{\Phi\left(z_1^i\right)-\Phi\left(z_2^i\right)}\right]^2 \quad (30)$$

Substituting Eq. (30) in Eq. (26) we obtain,

$$\Lambda_{ii} = \frac{1}{\sigma_\varepsilon^2}\left[\frac{z_1^i\mathcal{N}(z_1^i,0,1)-z_2^i\mathcal{N}(z_2^i,0,1)}{\Phi\left(z_1^i\right)-\Phi\left(z_2^i\right)}\right] + + \frac{1}{\sigma_\varepsilon^2}\left[\frac{\mathcal{N}(z_1^i,0,1)-\mathcal{N}(z_2^i,0,1)}{\Phi\left(z_1^i\right)-\Phi\left(z_2^i\right)}\right]^2 \quad (31)$$

We further note that the first term in Eq. (31) is positive. The constraint $\eta_{y_i} > \eta_{y_i-1}$ impose $z_1^i > z_2^i$ and therefore $\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right) > 0$. Additionally we can show that $z_1^i\mathcal{N}(z_1^i,0,1) - z_2^i\mathcal{N}(z_2^i,0,1) > 0$ (TBD). Hence, Λ is a positive-define matrix and hence H (24) is also positive definite (sum of positive definite matrices).

Appendix B

Evidence maximization is equivalent to finding the minimizer for the negative log evidence which according to (20) can be written as

$$g(\boldsymbol{\theta}) \triangleq -\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{i=1}^{n}\ell(y_i,\hat{f}_i,\boldsymbol{\theta}) + \frac{1}{2}\hat{\boldsymbol{f}}^T K^{-1}\hat{\boldsymbol{f}} + \frac{1}{2}\log|I + K\Lambda_{\mathrm{MAP}}|. \quad (32)$$

The hyper-parameter vector $\boldsymbol{\theta}$ include the threshold parameters $\{\eta_1, \log\Delta_2, ..., \log\Delta_{K-1}\}$, the noise variance $\log\sigma_\epsilon$ and the kernel parameters. The loss function ℓ and the covariance matrix K are function of $\boldsymbol{\theta}$, but $\hat{\boldsymbol{f}}$ and therefore Λ_{MAP} are also implicitly functions of $\boldsymbol{\theta}$, since when $\boldsymbol{\theta}$ changes, the optimum of the posterior $\hat{\boldsymbol{f}}$also changes. Thus

$$\frac{\partial g(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\partial g(\boldsymbol{\theta})}{\partial\theta_i}|_{\mathrm{explicit}} + \sum_{i=1}^{n}\frac{\partial g(\boldsymbol{\theta})}{\partial\hat{f}_i}\frac{\partial\hat{f}_i}{\partial\theta_j} \quad (33)$$

by the chain rule. The explicit term is given by

$$
\begin{aligned}
\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_j}|_{\text{explicit}} &= \frac{\partial}{\partial \theta_j}\sum_{i=1}^{n} \ell(y_i,\hat{f}_i,\boldsymbol{\theta}) - \frac{1}{2}\hat{\boldsymbol{f}}^T K^{-1}\frac{\partial K}{\partial \theta_j}K^{-1}\hat{\boldsymbol{f}} + \\
&+ \frac{1}{2}\text{trace}\left[(I+K\Lambda_{\text{MAP}})^{-1}\frac{\partial}{\partial \theta_j}K\Lambda_{\text{MAP}}\right] \\
&= \frac{\partial}{\partial \theta_j}\sum_{i=1}^{n} \ell(y_i,\hat{f}_i,\boldsymbol{\theta}) - \frac{1}{2}\hat{\boldsymbol{f}}^T K^{-1}\frac{\partial K}{\partial \theta_j}K^{-1}\hat{\boldsymbol{f}} \\
&+ \frac{1}{2}\text{trace}\left[(\Lambda_{\text{MAP}}^{-1}+K)^{-1}\frac{\partial K}{\partial \theta_j}\right] \\
&+ \frac{1}{2}\text{trace}\left[\Lambda_{\text{MAP}}^{-1}(\Lambda_{\text{MAP}}^{-1}+K)^{-1}K\frac{\partial}{\partial \theta_j}\frac{\partial^2 \ell(y_i,\hat{f}_i,\boldsymbol{\theta})}{\partial f_i^2}\right] =
\end{aligned}
\tag{34}
$$

where we used Eq. (19) for the expression of Λ_{MAP}. When evaluating the remaining term from Eq. (33), we utilize the fact that $\hat{\boldsymbol{f}}$ is the maximum of the posterior so that $\partial\Psi(\boldsymbol{f})/\partial\boldsymbol{f} = \boldsymbol{0}$ for $\boldsymbol{f} = \hat{\boldsymbol{f}}$ where $\Psi(\boldsymbol{f})$ is defined in Eq. (16). Thus the implicit derivatives of the first two term in Eq. (32) vanish, leaving only

$$
\begin{aligned}
\frac{\partial g(\boldsymbol{\theta})}{\partial \hat{f}_i} &= \frac{1}{2}\frac{\partial \log|I+K\Lambda_{\text{MAP}}|}{\partial \hat{f}_i} = \frac{1}{2}\text{trace}\left[(I+K\Lambda_{\text{MAP}})^{-1}K\frac{\partial \Lambda_{\text{MAP}}}{\partial \hat{f}_i}\right] \\
&= \frac{1}{2}\text{trace}\left[(K^{-1}+\Lambda_{\text{MAP}})^{-1}\frac{\partial \Lambda_{\text{MAP}}}{\partial \hat{f}_i}\right] = \frac{1}{2}\left[(K^{-1}+\Lambda_{\text{MAP}})^{-1}\right]_{ii}\frac{\partial^3 \ell(y_i\hat{f}_i,\boldsymbol{\theta})}{\partial f_i^3}
\end{aligned}
\tag{35}
$$

In order to evaluate the derivative $\partial\hat{\boldsymbol{f}}/\partial\theta_j$, we differentiate the fixed-point equation $\hat{\boldsymbol{f}} = -K\sum_{i=1}^{n}\partial\ell(y_i,f_i,\boldsymbol{\theta})/\partial\boldsymbol{f}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}}$ (Rasmussen et al.) to obtain

$$
\begin{aligned}
\frac{\partial \hat{\boldsymbol{f}}}{\partial \theta_j} &= -\frac{\partial K}{\partial \theta_j}\sum_{i=1}^{n}\frac{\partial \ell(y_i,f_i,\boldsymbol{\theta})}{\partial \boldsymbol{f}}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}} - K\frac{\partial}{\partial \theta_j}\sum_{i=1}^{n}\frac{\partial \ell(y_i,f_i,\boldsymbol{\theta})}{\partial f_i}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}} \\
&= -\frac{\partial K}{\partial \theta_j}\sum_{i=1}^{n}\frac{\partial \ell(y_i,f_i,\boldsymbol{\theta})}{\partial \boldsymbol{f}}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}} - K\sum_{i=1}^{n}\frac{\partial}{\partial \theta_j}\left[\frac{\partial \ell(y_i,f_i,\boldsymbol{\theta})}{\partial f_i}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}}\right] \\
&- K\underbrace{\frac{\partial}{\partial \hat{\boldsymbol{f}}}\sum_{i=1}^{n}\frac{\partial \ell(y_i,f_i,\boldsymbol{\theta})}{\partial f_i}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}}}_{\Lambda_{\text{MAP}}}\frac{\partial \hat{\boldsymbol{f}}}{\partial \theta_j} = -\frac{\partial K}{\partial \theta_j}\sum_{i=1}^{n}\frac{\partial \ell(y_i,f_i,\boldsymbol{\theta})}{\partial f_i}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}} \\
&- K\sum_{i=1}^{n}\frac{\partial}{\partial \theta_j}\left[\frac{\partial \ell(y_i,f_i,\boldsymbol{\theta})}{\partial f_i}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}}\right] - K\Lambda_{\text{MAP}}\frac{\partial \hat{\boldsymbol{f}}}{\partial \theta_j}
\end{aligned}
$$

Thus we obtain

$$\frac{\partial \hat{\boldsymbol{f}}}{\partial \theta_j} = -(I + K\Lambda_{\text{MAP}})^{-1} \left\{ \frac{\partial K}{\partial \theta_j} \sum_{i=1}^{n} \frac{\partial \ell(y_i, f_i, \boldsymbol{\theta})}{\partial f_i}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}} + K \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \left[\frac{\partial \ell(y_i, f_i, \boldsymbol{\theta})}{\partial f_i}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}} \right] \right\}$$

$$\begin{aligned}
\frac{\partial \hat{\boldsymbol{f}}}{\partial \theta_j} &= -\frac{\partial K}{\partial \theta_j} \sum_{i=1}^{n} \frac{\partial \ell(y_i, f_i, \boldsymbol{\theta})}{\partial f_i}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}} - K \frac{\partial}{\partial \hat{\boldsymbol{f}}} \sum_{i=1}^{n} \frac{\partial \ell(y_i, f_i, \boldsymbol{\theta})}{\partial f_i}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}} \frac{\partial \hat{\boldsymbol{f}}}{\partial \theta_j} \\
&= -\frac{\partial K}{\partial \theta_j} \sum_{i=1}^{n} \frac{\partial \ell(y_i, f_i, \boldsymbol{\theta})}{\partial f_i}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}} - K\Lambda_{\text{MAP}} \frac{\partial \hat{\boldsymbol{f}}}{\partial \theta_j} \\
&= -(I + K\Lambda_{\text{MAP}})^{-1} \frac{\partial K}{\partial \theta_j} \sum_{i=1}^{n} \frac{\partial \ell(y_i, f_i, \boldsymbol{\theta})}{\partial f_i}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}} \qquad (36)
\end{aligned}$$

The desired derivatives are obtained by plugging Eq. (34-36) to Eq. (33). We Note that the follwoing gradients are required for solving the optimization problem

$$\frac{\partial K}{\partial \theta_j}, \frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \ell(y_i, \hat{f}_i, \boldsymbol{\theta}), \frac{\partial}{\partial \theta_j} \frac{\partial \ell(y_i, \hat{f}_i, \boldsymbol{\theta})}{\partial f_i}, \frac{\partial}{\partial \theta_j} \frac{\partial^2 \ell(y_i, \hat{f}_i, \boldsymbol{\theta})}{\partial f_i^2}, \frac{\partial^3 \ell(y_i \hat{f}_i, \boldsymbol{\theta})}{\partial f_i^3}$$

Using Eq. (26) The derivative $\partial^3 \ell(y_i \hat{f}_i, \boldsymbol{\theta}) / \partial f_i^3$ in (35) becomes

$$\frac{\partial^3 \ell(y_i \hat{f}_i, \boldsymbol{\theta})}{\partial f_i^3} = -\frac{\partial^3}{\partial f_i^3} \log \left[\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right) \right] \qquad (37)$$

where $z_1^i = \frac{\eta_{y_i} - f_i}{\sigma_\epsilon}$ and $z_2^i = \frac{\eta_{y_i - 1} - f_i}{\sigma_\epsilon}$. Using the result in (30) we obtain

$$\frac{\partial^3 \ell(y_i \hat{f}_i, \boldsymbol{\theta})}{\partial f_i^3} = \frac{\partial}{\partial f_i} \frac{1}{\sigma_\epsilon^2} \left[v_0^2 + v_1 \right] = \frac{1}{\sigma_\epsilon^2} \left(2v_0 \frac{\partial v_0}{\partial f_i} + \frac{\partial v_1}{\partial f_i} \right) \qquad (38)$$

where

$$v_p^i = \frac{(z_1^i)^p \mathcal{N}(z_1^i, 0, 1) - (z_2^i)^p \mathcal{N}(z_2^i, 0, 1)}{\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)} \qquad (39)$$

Using Eq. (30) The term $\partial v_0 / \partial f_i$ can be expressed as

$$\frac{\partial v_0^i}{\partial f_i} = \frac{1}{\sigma_\epsilon} [(v_0^i)^2 + v_1^i] \qquad (40)$$

The term $\partial v_1 / \partial f_i$ can be expressed as

$$\frac{\partial v_1^i}{\partial f_i} = \frac{h'(f_i)k(f_i) - h(f_i)k'(f_i)}{k(f_i)^2} \qquad (41)$$

where

$$
\begin{aligned}
h(f_i) &= z_1^i\mathcal{N}(z_1^i,0,1) - z_2^i\mathcal{N}(z_2^i,0,1)\\
h'(f_i) &= \mathcal{N}(z_1^i,0,1)\frac{\partial z_1^i}{\partial f_i} - (z_1^i)^2\mathcal{N}(z_1^i,0,1)\frac{\partial z_1^i}{\partial f_i} - \mathcal{N}(z_2^i,0,1)\frac{\partial z_2^i}{\partial f_i} + (z_2^i)^2\mathcal{N}(z_2^i,0,1)\frac{\partial z_2^i}{\partial f_i}\\
&= -\frac{1}{\sigma_\epsilon}\mathcal{N}(z_1^i,0,1) + \frac{1}{\sigma_\epsilon}(z_1^i)^2\mathcal{N}(z_1^i,0,1) + \frac{1}{\sigma_\epsilon}\mathcal{N}(z_2^i,0,1) - \frac{1}{\sigma_\epsilon}(z_2^i)^2\mathcal{N}(z_2^i,0,1)\\
&= -\frac{1}{\sigma_\epsilon}\left[\mathcal{N}(z_1^i,0,1) - \mathcal{N}(z_2^i,0,1)\right] + \frac{1}{\sigma_\epsilon}\left[(z_1^i)^2\mathcal{N}(z_1^i,0,1) - (z_2^i)^2\mathcal{N}(z_2^i,0,1)\right]\\
k(f_i) &= \Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)\\
k'(f_i) &= \mathcal{N}(z_1^i,0,1)\frac{\partial z_1^i}{\partial f_i} - \mathcal{N}(z_2^i,0,1)\frac{\partial z_2^i}{\partial f_i}\\
&= -\frac{1}{\sigma_\epsilon}\left[\mathcal{N}(z_1^i,0,1) - \mathcal{N}(z_2^i,0,1)\right]
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\frac{\partial v_1}{\partial f_i} &= -\frac{1}{\sigma_\epsilon}\frac{\mathcal{N}(z_1^i,0,1) - \mathcal{N}(z_2^i,0,1)}{\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)} + \frac{1}{\sigma_\epsilon}\frac{(z_1^i)^2\mathcal{N}(z_1^i,0,1) - (z_2^i)^2\mathcal{N}(z_2^i,0,1)}{\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)}\\
&+ \frac{1}{\sigma_\epsilon}\frac{z_1^i\mathcal{N}(z_1^i,0,1) - z_2^i\mathcal{N}(z_2^i,0,1)}{\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)}\frac{\mathcal{N}(z_1^i,0,1) - \mathcal{N}(z_2^i,0,1)}{\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)}\\
&= \frac{1}{\sigma_\epsilon}\left[-v_0^i + v_2^i + v_1^i v_0^i\right] \qquad (42)
\end{aligned}
$$

Substituting Eq. (40,42) to Eq. (38) we obtain

$$
\frac{\partial^3\ell(y_i\hat{f}_i,\boldsymbol{\theta})}{\partial f_i^3} = \frac{1}{\sigma_\epsilon^2}\left(2v_0^i\frac{\partial v_0^i}{\partial f_i} + \frac{\partial v_1^i}{\partial f_i}\right) = \frac{1}{\sigma_\epsilon^3}[2(v_0^i)^3 + 3v_0^i v_1^i + v_2^i - v_0^i] \tag{43}
$$

The derivative of $\sum_{i=1}^n \ell(y_i,\hat{f}_i,\boldsymbol{\theta})$ with respect to θ_j is given by

$$
\frac{\partial}{\partial\theta_j}\sum_{i=1}^n \ell(y_i,\hat{f}_i,\boldsymbol{\theta}) = -\frac{\partial}{\partial\theta_j}\sum_{i=1}^n \log\left[\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)\right] = -\sum_{i=1}^n\frac{\mathcal{N}(z_1^i,0,1)\frac{\partial z_1^i}{\partial\theta_j} - \mathcal{N}(z_2^i,0,1)\frac{\partial z_2^i}{\partial\theta_j}}{\left[\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)\right]} \tag{44}
$$

The derivative of $\partial\ell(y_i\hat{f}_i,\boldsymbol{\theta})/\partial f_i$ and $\partial^2\ell(y_i,\hat{f}_i,\boldsymbol{\theta})/\partial f_i^2$ with respect to θ_j is given by (using Eq. 39)

$$
\frac{\partial}{\partial\theta_j}\frac{\partial\ell(y_i,\hat{f}_i,\boldsymbol{\theta})}{\partial f_i} = \frac{\partial}{\partial\theta_j}\frac{1}{\sigma_\epsilon}v_0^i \tag{45}
$$

$$
\frac{\partial}{\partial\theta_j}\frac{\partial^2\ell(y_i,\hat{f}_i,\boldsymbol{\theta})}{\partial f_i^2} = \frac{\partial}{\partial\theta_j}\frac{1}{\sigma_\epsilon^2}[(v_0^i)^2 + v_1^i] = \frac{1}{\sigma_\epsilon^2}\left(2v_0^i\frac{\partial v_0^i}{\partial\theta_j} + \frac{\partial v_1^i}{\partial\theta_j}\right) \tag{46}
$$

The derivatives with respect to σ_ϵ

For $\theta_i = \sigma_\epsilon$ we obtain

$$\frac{\partial z_1^i}{\partial \sigma_\epsilon} = -\frac{\eta_{y_i} - f_i}{\sigma_\epsilon^2} = -\frac{z_1^i}{\sigma_\epsilon}, \quad \frac{\partial z_2^i}{\partial \sigma_\epsilon} = -\frac{z_2^i}{\sigma_\epsilon}$$

The derivative of $\sum_{i=1}^n \ell(y_i, \hat{f}_i, \boldsymbol{\theta})$ with respect to σ_ϵ is given by

$$\frac{\partial}{\partial \sigma_\epsilon} \sum_{i=1}^n \ell(y_i, \hat{f}_i, \boldsymbol{\theta}) = \frac{1}{\sigma_\epsilon} \sum_{i=1}^n v_1^i \tag{47}$$

To compute the derivativein Eq. (45),(46) with respect to $\theta_j = \sigma_\epsilon$ we first compute the derivatives $\partial v_0 / \partial \sigma_\epsilon$ and $\partial v_1 / \partial \sigma_\epsilon$

$$\begin{aligned}
\frac{\partial v_0}{\partial \sigma_\epsilon} &= \frac{1}{\sigma_\epsilon} \frac{\left[(z_1^i)^2 \mathcal{N}(z_1^i, 0, 1) - (z_2^i)^2 \mathcal{N}(z_2^i, 0, 1)\right]}{\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)} \\
&= +\frac{1}{\sigma_\epsilon} \frac{\mathcal{N}(z_1^i, 0, 1) - \mathcal{N}(z_2^i, 0, 1)}{\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)} \frac{z_1^i \mathcal{N}(z_1^i, 0, 1) - z_2^i \mathcal{N}(z_2^i, 0, 1)}{\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)} \frac{\partial}{\partial \theta_j} \frac{\partial^2 \ell(y_i, \hat{f}_i, \boldsymbol{\theta})}{\partial f_i^2} \\
&= \frac{1}{\sigma_\epsilon} (v_2^i + v_0^i v_1^i)
\end{aligned} \tag{48}$$

$$\begin{aligned}
\frac{\partial v_1}{\partial \sigma_\epsilon} &= -\frac{1}{\sigma_\epsilon} \frac{z_1^i \mathcal{N}(z_1^i, 0, 1) - z_2^i \mathcal{N}(z_2^i, 0, 1)}{\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)} \\
&\quad + \frac{1}{\sigma_\epsilon} \frac{(z_1^i)^3 \mathcal{N}(z_1^i, 0, 1) - (z_2^i)^3 \mathcal{N}(z_2^i, 0, 1)}{\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)} \\
&\quad + \frac{1}{\sigma_\epsilon} \left(\frac{z_1^i \mathcal{N}(z_1^i, 0, 1) - z_2^i \mathcal{N}(z_2^i, 0, 1)}{\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)} \right)^2 \\
&= \frac{1}{\sigma_\epsilon} [-v_1^i + v_2^i + (v_1^i)^2]
\end{aligned}$$

Thus we obtain,

$$\frac{\partial}{\partial \sigma_\epsilon} \frac{\partial \ell(y_i, \hat{f}_i, \boldsymbol{\theta})}{\partial f_i} = \frac{1}{\sigma_\epsilon^2} (v_2^i + v_0^i v_1^i) \tag{49}$$

$$\frac{\partial}{\partial \sigma_\epsilon} \frac{\partial^2 \ell(y_i, \hat{f}_i, \boldsymbol{\theta})}{\partial f_i^2} = \frac{1}{\sigma_\epsilon^3} [2 v_0^i v_2^i + 2(v_0^i)^2 v_1^i - v_1^i + v_3^i + (v_1^i)^2] \tag{50}$$

The derivatives with respect to η_1

We note that

$$\eta_i = \begin{cases} -\infty & i = 0 \\ \eta_1 & i = 1 \\ \eta_1 + \sum_{l=1}^i \Delta_l & i \in [2, K-1] \\ \infty & i = K \end{cases} \tag{51}$$

For $\theta_j = \eta_1$ we obtain,

$$\frac{\partial z_1^i}{\partial \eta_1} = \frac{\partial}{\partial \eta_1} \frac{\eta_{y_i} - f_i}{\sigma_\epsilon} = \begin{cases} \frac{1}{\sigma_\epsilon} & y_i = [1, K-1] \\ 0 & y_i = K \end{cases}$$

$$\frac{\partial z_2^i}{\partial \eta_1} = \frac{\partial}{\partial \eta_1} \frac{\eta_{y_i - 1} - f_i}{\sigma_\epsilon} = \begin{cases} 0 & y_i = 1 \\ \frac{1}{\sigma_\epsilon} & y_i = [2, K] \end{cases}$$

We note however that

$$\mathcal{N}(z_1^i, 0, 1) \frac{\partial z_1^i}{\partial f_i} = \frac{1}{\sigma_\epsilon} \mathcal{N}(z_1^i, 0, 1)$$

$$\mathcal{N}(z_2^i, 0, 1) \frac{\partial z_2^i}{\partial f_i} = \frac{1}{\sigma_\epsilon} \mathcal{N}(z_2^i, 0, 1)$$

The derivative of $\sum_{i=1}^n \ell(y_i, \hat{f}_i, \boldsymbol{\theta})$ with respect to η_1 is given by (using Eq. 44)

$$\frac{\partial}{\partial \eta_1} \ell(y_i, \hat{f}_i, \boldsymbol{\theta}) = -\frac{1}{\sigma_\epsilon} v_0^i \tag{52}$$

To compute the derivativein Eq. (45),(46) with respect to $\theta_j = \eta_1$ we first compute the derivatives $\partial v_0 / \partial \eta_1$ and $\partial v_1 / \partial \eta_1$

$$\frac{\partial v_0^i}{\partial \eta_1} = -\frac{1}{\sigma_\epsilon} [(v_0^i)^2 + v_1^i]$$

$$\frac{\partial v_1^i}{\partial \eta_1} = -\frac{1}{\sigma_\epsilon} (-v_0^i + v_2^i + v_1^i v_0^i)$$

Thus we obtain,

$$\frac{\partial}{\partial \eta_1} \frac{\partial \ell(y_i, \hat{f}_i, \boldsymbol{\theta})}{\partial f_i} = -\frac{1}{\sigma_\epsilon^2} [(v_0)^2 + v_1^i] \tag{53}$$

$$\frac{\partial}{\partial \eta_1} \frac{\partial^2 \ell(y_i, \hat{f}_i, \boldsymbol{\theta})}{\partial f_i^2} = -\frac{1}{\sigma_\epsilon^3} \left[2(v_0^i)^3 + 3 v_0^i v_1 - v_0^i + v_2^i \right] \tag{54}$$

The derivatives with respect to Δ_l

Using Eq. (51) we obtain

$$\frac{\partial z_1^i}{\partial \Delta_l} = \begin{cases} 0 & y_i \leq l \\ \frac{1}{\sigma_\epsilon} & \text{else} \end{cases}, \quad l = 2, ..., K-1$$

$$\frac{\partial z_2^i}{\partial \Delta_l} = \begin{cases} 0 & y_i - 1 \leq l \\ \frac{1}{\sigma_\epsilon} & \text{else} \end{cases}, \quad l = 2, ..., K-1$$

The derivative of $\sum_{i=1}^{n} \ell(y_i, \hat{f}_i, \boldsymbol{\theta})$ with respect to Δ_l is given by (using Eq. 44)

$$\frac{\partial}{\partial \Delta_l} \ell(y_i, \hat{f}_i, \boldsymbol{\theta}) = \begin{cases} 0 & y_i < l \\ -\frac{1}{\sigma_\epsilon} s_0^i & y_i = l \;, \quad l = 2, ..., K-1 \\ -\frac{1}{\sigma_\epsilon} v_0^i & y_i > l \end{cases}$$

where

$$s_p = \frac{(z_1^i)^p \mathcal{N}(z_1^i, 0, 1)}{\Phi\left(z_1^i\right) - \Phi\left(z_2^i\right)} \tag{55}$$

To compute the derivativein Eq. (45),(46) with respect to $\theta_j = \eta_1$ we first compute the derivatives $\partial v_0^i / \partial \eta_1$ and $\partial v_1^i / \partial \eta_1$

$$\frac{\partial v_0^i}{\partial \Delta_l} = \begin{cases} 0 & y_i < l \\ -\frac{1}{\sigma_\epsilon}[s_1^i + v_0^i s_0^i] & y_i = l \;, \quad l = 2, ..., K-1 \\ -\frac{1}{\sigma_\epsilon}[(v_0^i)^2 + v_1^i] & y_i > l \end{cases}$$

$$\frac{\partial v_1^i}{\partial \Delta_l} = \begin{cases} 0 & y_i < l \\ -\frac{1}{\sigma_\epsilon}(-s_0^i + s_2^i + v_1^i s_0^i) & y_i = l \;, \quad l = 2, ..., K-1 \\ -\frac{1}{\sigma_\epsilon}(-v_0^i + v_2^i + v_1^i v_0^i) & y_i > l \end{cases}$$

Thus we obtain,

$$\frac{\partial}{\partial \Delta_l} \frac{\partial \ell(y_i, \hat{f}_i, \boldsymbol{\theta})}{\partial f_i} = \begin{cases} 0 & y_i < l \\ -\frac{1}{\sigma_\epsilon^2}[v_0^i s_0^i + s_1^i] & y_i = l \;, \quad l = 2, ..., K-1 \\ -\frac{1}{\sigma_\epsilon^2}[(v_0^i)^2 + v_1^i] & y_i > l \end{cases}$$

$$\frac{\partial}{\partial \Delta_l} \frac{\partial^2 \ell(y_i, \hat{f}_i, \boldsymbol{\theta})}{\partial f_i^2} = \begin{cases} 0 & y_i < l \\ -\frac{1}{\sigma_\epsilon^3}\left[2(v_0^i)^2 s_0^i + 2 v_0^i s_1^i - s_0^i + s_2^i + v_1^i s_0^i\right] & y_i = l \\ -\frac{1}{\sigma_\epsilon^3}\left[2(v_0^i)^3 + 3 v_0^i v_1^i - v_0^i + v_2^i\right] & y_i > l \end{cases}$$

References

[1] Chu, Wei, Zoubin Ghahramani, and Christopher KI Williams. "Gaussian processes for ordinal regression." Journal of machine learning research 6.7 (2005).