

# Dynamic ordinal regression modeling, with applications to estimating natural selection surfaces

Athanasios Kottas

*Department of Applied Mathematics and Statistics, University of California, Santa Cruz*

Joint work with Maria DeYoreo (*RAND Corporation*)

BIRS Workshop: Bayesian Nonparametric Inference: Dependence Structures and their Applications  
Oaxaca, Mexico – December 3-8, 2017

# Motivation

- Regression modeling for one or more **ordinal categorical responses** recorded over discrete time
- Focus on applications, including problems in **ecology** and the **environmental sciences**, where it is natural/necessary to model the joint stochastic mechanism for the response(s) and covariates
- Motivating application: study of dynamically evolving natural selection surfaces in evolutionary/population biology
- Data example: maturity (recorded on an ordinal scale), length, and age for Chilipepper rockfish, collected over 15 years along the coast of California

# Motivation

- Regression modeling for one or more **ordinal categorical responses** recorded over discrete time
- Focus on applications, including problems in **ecology** and the **environmental sciences**, where it is natural/necessary to model the joint stochastic mechanism for the response(s) and covariates
- Motivating application: study of dynamically evolving natural selection surfaces in evolutionary/population biology
- Data example: maturity (recorded on an ordinal scale), length, and age for Chilipepper rockfish, collected over 15 years along the coast of California

# Motivation

- Regression modeling for one or more **ordinal categorical responses** recorded over discrete time
- Focus on applications, including problems in **ecology** and the **environmental sciences**, where it is natural/necessary to model the joint stochastic mechanism for the response(s) and covariates
- Motivating application: study of dynamically evolving natural selection surfaces in evolutionary/population biology
- Data example: maturity (recorded on an ordinal scale), length, and age for Chilipepper rockfish, collected over 15 years along the coast of California

# Modeling through latent continuous responses

- Assume each ordinal response represents a discretized version of an underlying **latent continuous response**<sup>1</sup>
- $k$  ordinal variables  $Y = (Y_1, \dots, Y_k)$ , with  $y_j \in \{1, \dots, C_j\}$ , and  $p$  (continuous) covariates  $X = (X_1, \dots, X_p)$
- Assume  $Y_j = l$  if-f  $\gamma_{j,l-1} < Z_j \leq \gamma_{j,l}$ , for  $j = 1, \dots, k$ , and  $l = 1, \dots, C_j$  (with  $\gamma_{j,0} = -\infty$  and  $\gamma_{j,C_j} = \infty$ )
- Multivariate normal distribution for  $Z = (Z_1, \dots, Z_k) \rightarrow$  multivariate ordinal probit model
  - ▷ symmetric, unimodal latent response distribution with mean  $x^T \beta \rightarrow$  implies restrictive effects of covariates on the probability response curves
  - ▷ computational challenges in estimating cut-off points

---

<sup>1</sup>e.g., Albert and Chib, 1993

# Modeling through latent continuous responses

- Assume each ordinal response represents a discretized version of an underlying **latent continuous response**<sup>1</sup>
- $k$  ordinal variables  $\mathbf{Y} = (Y_1, \dots, Y_k)$ , with  $y_j \in \{1, \dots, C_j\}$ , and  $p$  (continuous) covariates  $\mathbf{X} = (X_1, \dots, X_p)$
- Assume  $Y_j = l$  if-f  $\gamma_{j,l-1} < Z_j \leq \gamma_{j,l}$ , for  $j = 1, \dots, k$ , and  $l = 1, \dots, C_j$  (with  $\gamma_{j,0} = -\infty$  and  $\gamma_{j,C_j} = \infty$ )
- Multivariate normal distribution for  $\mathbf{Z} = (Z_1, \dots, Z_k) \rightarrow$  multivariate ordinal probit model
  - ▷ symmetric, unimodal latent response distribution with mean  $\mathbf{x}^T \boldsymbol{\beta} \rightarrow$  implies restrictive effects of covariates on the probability response curves
  - ▷ computational challenges in estimating cut-off points

<sup>1</sup>e.g., Albert and Chib, 1993

# Modeling through latent continuous responses

- Assume each ordinal response represents a discretized version of an underlying **latent continuous response**<sup>1</sup>
- $k$  ordinal variables  $\mathbf{Y} = (Y_1, \dots, Y_k)$ , with  $y_j \in \{1, \dots, C_j\}$ , and  $p$  (continuous) covariates  $\mathbf{X} = (X_1, \dots, X_p)$
- Assume  $Y_j = l$  if-f  $\gamma_{j,l-1} < Z_j \leq \gamma_{j,l}$ , for  $j = 1, \dots, k$ , and  $l = 1, \dots, C_j$  (with  $\gamma_{j,0} = -\infty$  and  $\gamma_{j,C_j} = \infty$ )
- Multivariate normal distribution for  $\mathbf{Z} = (Z_1, \dots, Z_k) \rightarrow$  multivariate ordinal probit model
  - ▷ symmetric, unimodal latent response distribution with mean  $\mathbf{x}^T \boldsymbol{\beta} \rightarrow$  implies restrictive effects of covariates on the probability response curves
  - ▷ computational challenges in estimating cut-off points

---

<sup>1</sup>e.g., Albert and Chib, 1993

# Objectives

- For univariate responses, more general methods have been explored, relaxing either the distributional or linearity assumption <sup>1</sup>
- In the multivariate setting, complications arise from issues of constrained covariance matrices and inference for the cut-offs, and methods for general Bayesian inference are limited
- In contrast to semiparametric approaches, our aim is flexible modeling and inference for the ordinal regression relationships **and** for the response distribution

---

<sup>1</sup>e.g., Newton et al., 1996; Mukhopadhyay and Gelfand, 1997; Denison et al., 2002; Chib and Greenberg, 2010

# Objectives

- For univariate responses, more general methods have been explored, relaxing either the distributional or linearity assumption <sup>1</sup>
- In the multivariate setting, complications arise from issues of constrained covariance matrices and inference for the cut-offs, and methods for general Bayesian inference are limited
- In contrast to semiparametric approaches, our aim is flexible modeling and inference for the ordinal regression relationships **and** for the response distribution

---

<sup>1</sup>e.g., Newton et al., 1996; Mukhopadhyay and Gelfand, 1997; Denison et al., 2002; Chib and Greenberg, 2010

# Objectives

- For univariate responses, more general methods have been explored, relaxing either the distributional or linearity assumption <sup>1</sup>
- In the multivariate setting, complications arise from issues of constrained covariance matrices and inference for the cut-offs, and methods for general Bayesian inference are limited
- In contrast to semiparametric approaches, our aim is flexible modeling and inference for the ordinal regression relationships **and** for the response distribution

---

<sup>1</sup>e.g., Newton et al., 1996; Mukhopadhyay and Gelfand, 1997; Denison et al., 2002; Chib and Greenberg, 2010

# The nonparametric mixture model

- We use a version of **implied conditional regression**<sup>1</sup> modeling the joint latent response-covariate distribution  $f(\mathbf{z}, \mathbf{x})$ 
  - ▷ inference for  $f(\mathbf{z} | \mathbf{x})$ , and for  $\Pr(\mathbf{Y} | \mathbf{x})$ , implied through  $f(\mathbf{z}, \mathbf{x})$  and  $f(\mathbf{x})$
- Dirichlet Process (DP) mixture model for  $f(\mathbf{z}, \mathbf{x})$ :

$$f(\mathbf{z}, \mathbf{x} | G) = \int \mathbf{N}(\mathbf{z}, \mathbf{x} | \boldsymbol{\mu}, \Sigma) dG(\boldsymbol{\mu}, \Sigma), \quad G | \alpha, \boldsymbol{\psi} \sim \text{DP}(\alpha, G_0(\cdot | \boldsymbol{\psi}))$$

- DP constructive definition<sup>2</sup>:  
 $f(\mathbf{z}, \mathbf{x} | G) = \sum_{r=1}^{\infty} p_r \mathbf{N}(\mathbf{z}, \mathbf{x} | \boldsymbol{\mu}_r, \Sigma_r)$ , where  $\boldsymbol{\theta}_r = (\boldsymbol{\mu}_r, \Sigma_r) \stackrel{iid}{\sim} G_0$ , and the weights  $p_1, p_2, \dots$  are determined through stick-breaking:
  - ▷ stick-breaking proportions  $\beta_s \stackrel{iid}{\sim} \text{beta}(1, \alpha)$ ,  $s = 1, 2, \dots$
  - ▷  $p_1 = \beta_1$ , and  $p_r = \beta_r \prod_{m=1}^{r-1} (1 - \beta_m)$ , for  $r = 2, 3, \dots$

---

<sup>1</sup>Nadaraya, 1964; Watson, 1964; Müller et al., 1996

<sup>2</sup>Sethuraman, 1994

# The nonparametric mixture model

- We use a version of **implied conditional regression**<sup>1</sup> modeling the joint latent response-covariate distribution  $f(\mathbf{z}, \mathbf{x})$ 
  - ▷ inference for  $f(\mathbf{z} | \mathbf{x})$ , and for  $\Pr(\mathbf{Y} | \mathbf{x})$ , implied through  $f(\mathbf{z}, \mathbf{x})$  and  $f(\mathbf{x})$
- Dirichlet Process (DP) mixture model for  $f(\mathbf{z}, \mathbf{x})$ :

$$f(\mathbf{z}, \mathbf{x} | G) = \int \mathbf{N}(\mathbf{z}, \mathbf{x} | \boldsymbol{\mu}, \Sigma) dG(\boldsymbol{\mu}, \Sigma), \quad G | \alpha, \boldsymbol{\psi} \sim \text{DP}(\alpha, G_0(\cdot | \boldsymbol{\psi}))$$

- DP constructive definition<sup>2</sup>:

$f(\mathbf{z}, \mathbf{x} | G) = \sum_{r=1}^{\infty} p_r \mathbf{N}(\mathbf{z}, \mathbf{x} | \boldsymbol{\mu}_r, \Sigma_r)$ , where  $\boldsymbol{\theta}_r = (\boldsymbol{\mu}_r, \Sigma_r) \stackrel{iid}{\sim} G_0$ , and the weights  $p_1, p_2, \dots$  are determined through stick-breaking:

- ▷ stick-breaking proportions  $\beta_s \stackrel{iid}{\sim} \text{beta}(1, \alpha)$ ,  $s = 1, 2, \dots$
- ▷  $p_1 = \beta_1$ , and  $p_r = \beta_r \prod_{m=1}^{r-1} (1 - \beta_m)$ , for  $r = 2, 3, \dots$

<sup>1</sup>Nadaraya, 1964; Watson, 1964; Müller et al., 1996

<sup>2</sup>Sethuraman, 1994

# The nonparametric mixture model

- We use a version of **implied conditional regression**<sup>1</sup> modeling the joint latent response-covariate distribution  $f(\mathbf{z}, \mathbf{x})$ 
  - ▷ inference for  $f(\mathbf{z} | \mathbf{x})$ , and for  $\Pr(\mathbf{Y} | \mathbf{x})$ , implied through  $f(\mathbf{z}, \mathbf{x})$  and  $f(\mathbf{x})$
- Dirichlet Process (DP) mixture model for  $f(\mathbf{z}, \mathbf{x})$ :

$$f(\mathbf{z}, \mathbf{x} | G) = \int \mathbf{N}(\mathbf{z}, \mathbf{x} | \boldsymbol{\mu}, \Sigma) dG(\boldsymbol{\mu}, \Sigma), \quad G | \alpha, \boldsymbol{\psi} \sim \text{DP}(\alpha, G_0(\cdot | \boldsymbol{\psi}))$$

- DP constructive definition<sup>2</sup>:  
 $f(\mathbf{z}, \mathbf{x} | G) = \sum_{r=1}^{\infty} p_r \mathbf{N}(\mathbf{z}, \mathbf{x} | \boldsymbol{\mu}_r, \Sigma_r)$ , where  $\boldsymbol{\theta}_r = (\boldsymbol{\mu}_r, \Sigma_r) \stackrel{iid}{\sim} G_0$ , and the weights  $p_1, p_2, \dots$  are determined through stick-breaking:
  - ▷ stick-breaking proportions  $\beta_s \stackrel{iid}{\sim} \text{beta}(1, \alpha)$ ,  $s = 1, 2, \dots$
  - ▷  $p_1 = \beta_1$ , and  $p_r = \beta_r \prod_{m=1}^{r-1} (1 - \beta_m)$ , for  $r = 2, 3, \dots$

---

<sup>1</sup>Nadaraya, 1964; Watson, 1964; Müller et al., 1996

<sup>2</sup>Sethuraman, 1994

# Ordinal regression functions

- Flexible model for  $f(\mathbf{z}, \mathbf{x}) \rightarrow$  flexible inference for  $\Pr(\mathbf{Y} | \mathbf{x})$
- Implied regression functions provide a nonparametric extension of probit regression (with random covariates):

$$\Pr(Y = (l_1, \dots, l_k) | \mathbf{x}; G) = \sum_{r=1}^{\infty} w_r(\mathbf{x}) \int_{\gamma_{k,l_k-1}}^{\gamma_{k,l_k}} \dots \int_{\gamma_{1,l_1-1}}^{\gamma_{1,l_1}} \mathbf{N}(\mathbf{z} | m_r(\mathbf{x}), S_r) d\mathbf{z}$$

▷ with covariate dependent weights  $w_r(\mathbf{x}) \propto p_r \mathbf{N}(\mathbf{x} | \boldsymbol{\mu}_r^x, \Sigma_r^{xx})$

▷ and covariate dependent probabilities, where

$$m_r(\mathbf{x}) = \boldsymbol{\mu}_r^z + \Sigma_r^{zx} (\Sigma_r^{xx})^{-1} (\mathbf{x} - \boldsymbol{\mu}_r^x) \text{ and } S_r = \Sigma_r^{zz} - \Sigma_r^{zx} (\Sigma_r^{xx})^{-1} \Sigma_r^{xz}$$

# Ordinal regression functions

- Flexible model for  $f(\mathbf{z}, \mathbf{x}) \rightarrow$  flexible inference for  $\Pr(\mathbf{Y} | \mathbf{x})$
- **Implied regression functions** provide a nonparametric extension of probit regression (with random covariates):

$$\Pr(\mathbf{Y} = (l_1, \dots, l_k) | \mathbf{x}; G) = \sum_{r=1}^{\infty} w_r(\mathbf{x}) \int_{\gamma^{k, l_k-1}}^{\gamma^{k, l_k}} \dots \int_{\gamma^{1, l_1-1}}^{\gamma^{1, l_1}} \mathbf{N}(\mathbf{z} | m_r(\mathbf{x}), S_r) d\mathbf{z}$$

▷ with covariate dependent weights  $w_r(\mathbf{x}) \propto p_r \mathbf{N}(\mathbf{x} | \boldsymbol{\mu}_r^x, \Sigma_r^{xx})$

▷ and covariate dependent probabilities, where

$$m_r(\mathbf{x}) = \boldsymbol{\mu}_r^z + \Sigma_r^{zx} (\Sigma_r^{xx})^{-1} (\mathbf{x} - \boldsymbol{\mu}_r^x) \text{ and } S_r = \Sigma_r^{zz} - \Sigma_r^{zx} (\Sigma_r^{xx})^{-1} \Sigma_r^{xz}$$

# Model properties

- Provided  $C_j > 2$ , both  $\mu$  and  $\Sigma$  are identifiable in the induced mixture kernel for  $(Y, X)$ , under **fixed cut-off points**
- The prior model has **large support** again under **fixed cut-offs**
  - ▷ it assigns positive probability to all Kullback-Leibler (KL) neighborhoods of a mixed ordinal-continuous distribution,  $p_0(x, y)$ , as well as to all KL neighborhoods of the implied conditional distribution,  $p_0(y | x)$
- KL property (and identifiability result) obtained under fixed cut-offs
  - **computational advantage** over parametric models
- More structured model for binary responses to incorporate identifiability restrictions for  $\Sigma$  (DeYoreo & Kottas, 2015, BA)

# Model properties

- Provided  $C_j > 2$ , both  $\mu$  and  $\Sigma$  are identifiable in the induced mixture kernel for  $(Y, X)$ , under **fixed cut-off points**
- The prior model has **large support** again under **fixed cut-offs**
  - ▷ it assigns positive probability to all Kullback-Leibler (KL) neighborhoods of a mixed ordinal-continuous distribution,  $p_0(\mathbf{x}, \mathbf{y})$ , as well as to all KL neighborhoods of the implied conditional distribution,  $p_0(\mathbf{y} \mid \mathbf{x})$
- KL property (and identifiability result) obtained under fixed cut-offs
  - **computational advantage** over parametric models
- More structured model for binary responses to incorporate identifiability restrictions for  $\Sigma$  (DeYoreo & Kottas, 2015, BA)

# Model properties

- Provided  $C_j > 2$ , both  $\mu$  and  $\Sigma$  are identifiable in the induced mixture kernel for  $(Y, X)$ , under **fixed cut-off points**
- The prior model has **large support** again under **fixed cut-offs**
  - ▷ it assigns positive probability to all Kullback-Leibler (KL) neighborhoods of a mixed ordinal-continuous distribution,  $p_0(\mathbf{x}, \mathbf{y})$ , as well as to all KL neighborhoods of the implied conditional distribution,  $p_0(\mathbf{y} \mid \mathbf{x})$
- KL property (and identifiability result) obtained under fixed cut-offs
  - **computational advantage** over parametric models
- More structured model for binary responses to incorporate identifiability restrictions for  $\Sigma$  (DeYoreo & Kottas, 2015, BA)

# Model properties

- Provided  $C_j > 2$ , both  $\mu$  and  $\Sigma$  are identifiable in the induced mixture kernel for  $(Y, X)$ , under **fixed cut-off points**
- The prior model has **large support** again under **fixed cut-offs**
  - ▷ it assigns positive probability to all Kullback-Leibler (KL) neighborhoods of a mixed ordinal-continuous distribution,  $p_0(\mathbf{x}, \mathbf{y})$ , as well as to all KL neighborhoods of the implied conditional distribution,  $p_0(\mathbf{y} \mid \mathbf{x})$
- KL property (and identifiability result) obtained under fixed cut-offs
  - **computational advantage** over parametric models
- More structured model for binary responses to incorporate identifiability restrictions for  $\Sigma$  (DeYoreo & Kottas, 2015, BA)

# Model properties (and implementation)

- Interactions and dependence between covariates are implicit in joint modeling framework
- Inference for inverse relationships  $\rightarrow$  covariate distribution across ordinal responses values,  $f(\mathbf{x} \mid \mathbf{Y} = \mathbf{y})$
- Model can accommodate directly continuous covariates as well as ordinal categorical covariates
- Posterior simulation: given the continuous latent responses, a normal DP mixture model (only extra step involves imputing the latent variables)

# Model properties (and implementation)

- Interactions and dependence between covariates are implicit in joint modeling framework
- Inference for inverse relationships  $\rightarrow$  covariate distribution across ordinal responses values,  $f(\mathbf{x} \mid \mathbf{Y} = \mathbf{y})$
- Model can accommodate directly continuous covariates as well as ordinal categorical covariates
- Posterior simulation: given the continuous latent responses, a normal DP mixture model (only extra step involves imputing the latent variables)

# Model properties (and implementation)

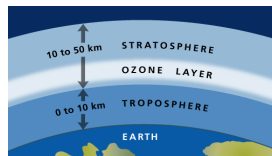
- **Interactions and dependence** between covariates are implicit in joint modeling framework
- Inference for **inverse relationships**  $\rightarrow$  covariate distribution across ordinal responses values,  $f(\mathbf{x} \mid \mathbf{Y} = \mathbf{y})$
- Model can accommodate directly continuous covariates as well as ordinal categorical covariates
- Posterior simulation: given the continuous latent responses, a normal DP mixture model (only extra step involves imputing the latent variables)

# Model properties (and implementation)

- **Interactions and dependence** between covariates are implicit in joint modeling framework
- Inference for **inverse relationships**  $\rightarrow$  covariate distribution across ordinal responses values,  $f(\mathbf{x} \mid \mathbf{Y} = \mathbf{y})$
- Model can accommodate directly continuous covariates as well as ordinal categorical covariates
- Posterior simulation: given the continuous latent responses, a normal DP mixture model (only extra step involves imputing the latent variables)

# Ozone concentration data example

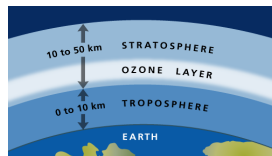
- Data set comprising 111 measurements of **ozone concentration** (ppb), wind speed (mph), radiation (langleys), and temperature (degrees Fahrenheit)



- Ozone concentration recorded on continuous scale
- To construct an ordinal response: define “high” as above 100 ppb, “medium” as  $(50, 100]$  ppb, and “low” as less than 50 ppb
- Comparison of inferences from the model for  $(Y, X)$  with those from a DP mixture of normals model for  $(Z, X)$

# Ozone concentration data example

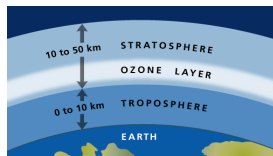
- Data set comprising 111 measurements of **ozone concentration** (ppb), wind speed (mph), radiation (langleys), and temperature (degrees Fahrenheit)



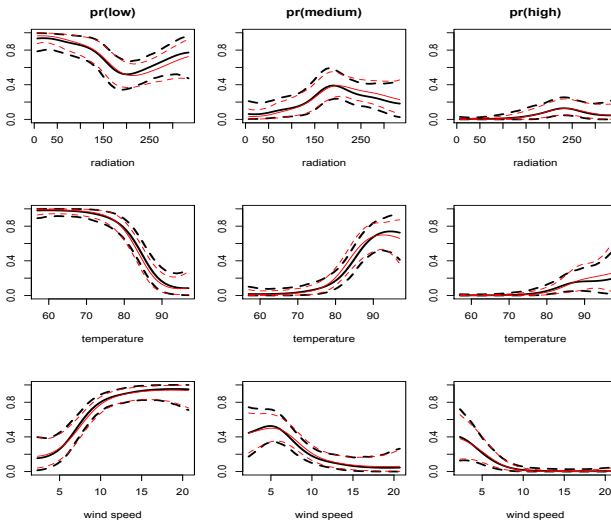
- Ozone concentration recorded on continuous scale
- To construct an ordinal response: define “high” as above 100 ppb, “medium” as (50, 100] ppb, and “low” as less than 50 ppb
- Comparison of inferences from the model for  $(Y, X)$  with those from a DP mixture of normals model for  $(Z, X)$

# Ozone concentration data example

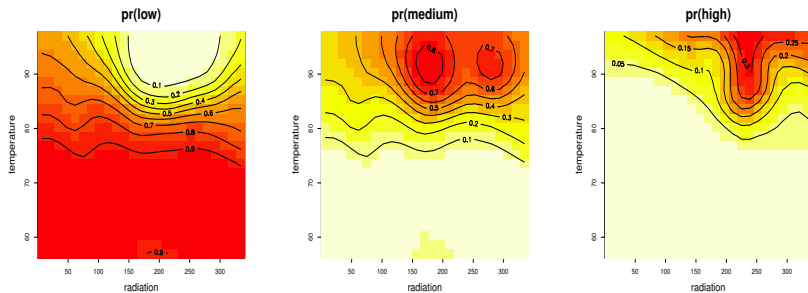
- Data set comprising 111 measurements of **ozone concentration** (ppb), wind speed (mph), radiation (langleys), and temperature (degrees Fahrenheit)



- Ozone concentration recorded on continuous scale
- To construct an ordinal response: define “high” as above 100 ppb, “medium” as (50, 100] ppb, and “low” as less than 50 ppb
- Comparison of inferences from the model for  $(Y, X)$  with those from a DP mixture of normals model for  $(Z, X)$



**Figure:** Posterior mean (solid) and 95% interval estimates (dashed) for  $\Pr(Y = l | x_m; G)$  (black) compared to  $\Pr(\gamma_{l-1} < Z \leq \gamma_l | x_m; G)$  (red).



**Figure:** Posterior mean estimates for  $\Pr(Y = l \mid x_1, x_2; G)$ , for  $l = 1, 2, 3$ , corresponding to low (left), medium (middle) and high (right). Red represents a value of 1, white represents 0.

# Extension to dynamic ordinal regression modeling

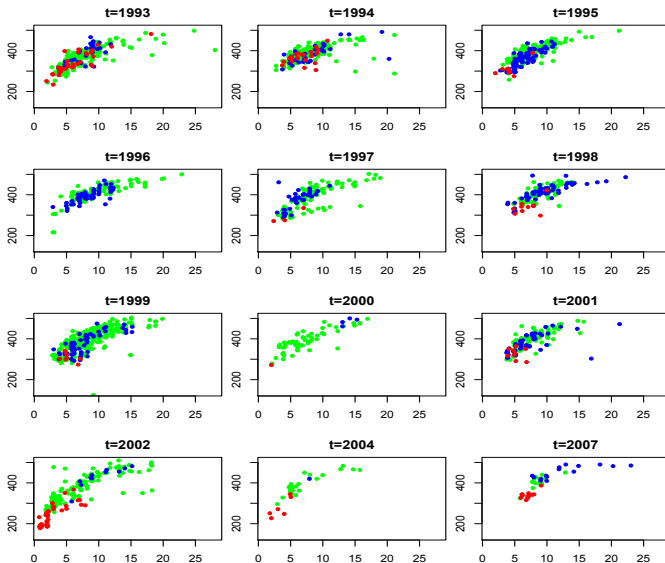
- Focusing on a univariate ordinal response, we seek to extend to a model for  $\Pr_t(Y | \mathbf{x})$ , for  $t \in \mathcal{T} = \{1, 2, \dots\}$
- Build on the earlier framework by extending to a prior model for  $\{f(z, \mathbf{x} | G_t) : t \in \mathcal{T}\}$ , and thus for  $\{\Pr(Y | \mathbf{x}, G_t) : t \in \mathcal{T}\}$
- Motivating application: data from NMFS on female Chilipepper rockfish collected between 1993 and 2007 along the coast of California
  - ▷ sample sizes per year range from 37 to 396, with no data available for three years (2003, 2005 and 2006)
  - ▷ three ordinal levels for maturity: immature (1), pre-spawning mature (2), and post-spawning mature (3)
  - ▷ length measured in millimeters
  - ▷ age recorded on an ordinal scale: age  $j$  implies the fish was between  $j$  and  $j + 1$  years of age (data range: 1 to 25) → incorporate age into the model in the same fashion with the maturity variable

# Extension to dynamic ordinal regression modeling

- Focusing on a univariate ordinal response, we seek to extend to a model for  $\Pr_t(Y | \mathbf{x})$ , for  $t \in \mathcal{T} = \{1, 2, \dots\}$
- Build on the earlier framework by extending to a prior model for  $\{f(\mathbf{z}, \mathbf{x} | G_t) : t \in \mathcal{T}\}$ , and thus for  $\{\Pr(Y | \mathbf{x}, G_t) : t \in \mathcal{T}\}$
- Motivating application: data from NMFS on female Chilipepper rockfish collected between 1993 and 2007 along the coast of California
  - ▷ sample sizes per year range from 37 to 396, with no data available for three years (2003, 2005 and 2006)
  - ▷ three ordinal levels for maturity: immature (1), pre-spawning mature (2), and post-spawning mature (3)
  - ▷ length measured in millimeters
  - ▷ age recorded on an ordinal scale: age  $j$  implies the fish was between  $j$  and  $j + 1$  years of age (data range: 1 to 25) → incorporate age into the model in the same fashion with the maturity variable

# Extension to dynamic ordinal regression modeling

- Focusing on a univariate ordinal response, we seek to extend to a model for  $\Pr_t(Y | \mathbf{x})$ , for  $t \in \mathcal{T} = \{1, 2, \dots\}$
- Build on the earlier framework by extending to a prior model for  $\{f(\mathbf{z}, \mathbf{x} | G_t) : t \in \mathcal{T}\}$ , and thus for  $\{\Pr(Y | \mathbf{x}, G_t) : t \in \mathcal{T}\}$
- Motivating application: data from NMFS on female Chilipepper rockfish collected between 1993 and 2007 along the coast of California
  - ▷ sample sizes per year range from 37 to 396, with no data available for three years (2003, 2005 and 2006)
  - ▷ three ordinal levels for maturity: immature (1), pre-spawning mature (2), and post-spawning mature (3)
  - ▷ length measured in millimeters
  - ▷ age recorded on an ordinal scale: age  $j$  implies the fish was between  $j$  and  $j + 1$  years of age (data range: 1 to 25) → incorporate age into the model in the same fashion with the maturity variable



**Figure:** Bivariate plots of length versus age at each year of data, with data points colored according to maturity level: red level 1; green level 2; blue level 3.

# DDP model extension

- To retain model properties at each  $t$ , use DDP prior for  $\{G_t : t \in \mathcal{T}\}$ <sup>1</sup>
- Time-dependent weights and atoms:

$$f(\mathbf{z}, \mathbf{x} \mid G_t) = \sum_{r=1}^{\infty} \left\{ (1 - \beta_{r,t}) \prod_{m=1}^{r-1} \beta_{m,t} \right\} \text{N}(\mathbf{z}, \mathbf{x} \mid \boldsymbol{\mu}_{r,t}, \boldsymbol{\Sigma}_r)$$

- Vector autoregressive model for the  $\{\boldsymbol{\mu}_{r,t} : t \in \mathcal{T}\}$ 
  - ▷  $\boldsymbol{\mu}_{r,t} \mid \boldsymbol{\mu}_{r,t-1}, \Theta, \mathbf{m}, \mathbf{V} \sim \text{N}(\mathbf{m} + \Theta \boldsymbol{\mu}_{r,t-1}, \mathbf{V})$
  - ▷  $\boldsymbol{\Sigma}_r \mid \nu, \mathbf{D} \stackrel{iid}{\sim} \text{IW}(\nu, \mathbf{D})$
  - ▷ hyperpriors for  $(\Theta, \mathbf{m}, \mathbf{V})$  and for  $\mathbf{D}$

---

<sup>1</sup>MacEachern, 2000

# DDP model extension

- To retain model properties at each  $t$ , use DDP prior for  $\{G_t : t \in \mathcal{T}\}$ <sup>1</sup>
- Time-dependent weights and atoms:

$$f(\mathbf{z}, \mathbf{x} \mid G_t) = \sum_{r=1}^{\infty} \left\{ (1 - \beta_{r,t}) \prod_{m=1}^{r-1} \beta_{m,t} \right\} \mathbf{N}(\mathbf{z}, \mathbf{x} \mid \boldsymbol{\mu}_{r,t}, \boldsymbol{\Sigma}_r)$$

- Vector autoregressive model for the  $\{\boldsymbol{\mu}_{r,t} : t \in \mathcal{T}\}$ 
  - ▷  $\boldsymbol{\mu}_{r,t} \mid \boldsymbol{\mu}_{r,t-1}, \Theta, \mathbf{m}, V \sim \mathbf{N}(\mathbf{m} + \Theta \boldsymbol{\mu}_{r,t-1}, V)$
  - ▷  $\boldsymbol{\Sigma}_r \mid \nu, \mathbf{D} \stackrel{iid}{\sim} \text{IW}(\nu, \mathbf{D})$
  - ▷ hyperpriors for  $(\Theta, \mathbf{m}, V)$  and for  $\mathbf{D}$

<sup>1</sup>MacEachern, 2000

# DDP model extension

- To retain model properties at each  $t$ , use DDP prior for  $\{G_t : t \in \mathcal{T}\}$ <sup>1</sup>
- Time-dependent weights and atoms:

$$f(\mathbf{z}, \mathbf{x} \mid G_t) = \sum_{r=1}^{\infty} \left\{ (1 - \beta_{r,t}) \prod_{m=1}^{r-1} \beta_{m,t} \right\} \mathbf{N}(\mathbf{z}, \mathbf{x} \mid \boldsymbol{\mu}_{r,t}, \boldsymbol{\Sigma}_r)$$

- Vector autoregressive model for the  $\{\boldsymbol{\mu}_{r,t} : t \in \mathcal{T}\}$ 
  - ▷  $\boldsymbol{\mu}_{r,t} \mid \boldsymbol{\mu}_{r,t-1}, \Theta, \mathbf{m}, \mathbf{V} \sim \mathbf{N}(\mathbf{m} + \Theta \boldsymbol{\mu}_{r,t-1}, \mathbf{V})$
  - ▷  $\boldsymbol{\Sigma}_r \mid \nu, \mathbf{D} \stackrel{iid}{\sim} \text{IW}(\nu, \mathbf{D})$
  - ▷ hyperpriors for  $(\Theta, \mathbf{m}, \mathbf{V})$  and for  $\mathbf{D}$

<sup>1</sup>MacEachern, 2000

# DDP model extension

- Stochastic process with  $\text{beta}(\alpha, 1)$  marginals:

$$\mathcal{B} = \left\{ \beta_t = \exp \left( -\frac{\zeta^2 + \eta_t^2}{2\alpha} \right) : t \in \mathcal{T} \right\}$$

where  $\zeta \sim N(0, 1)$  and, independently,  $\{\eta_t : t \in \mathcal{T}\}$  arises from a time series model with  $N(0, 1)$  marginals

- Build model for the  $\{\beta_{r,t} : t \in \mathcal{T}\}$  from  $\beta_{r,t} = \exp\{-(\zeta_r^2 + \eta_{r,t}^2)/(2\alpha)\}$ 
  - $\zeta_r \stackrel{\text{ind.}}{\sim} N(0, 1)$
  - AR(1) process for the  $\{\eta_{r,t} : t \in \mathcal{T}\}$ :  $\eta_{r,t} \mid \eta_{r,t-1}, \phi \sim N(\phi\eta_{r,t-1}, 1 - \phi^2)$  with  $|\phi| < 1$  (and  $\eta_{r,1} \stackrel{\text{ind.}}{\sim} N(0, 1)$ )
- Different types of correlations can be studied:  $\text{corr}(p_{r,t}, p_{r,t+k})$ , and  $\text{corr}(G_t(A), G_{t+1}(A))$ , for any subset  $A$  in the support of the  $G_t$

# DDP model extension

- Stochastic process with beta( $\alpha, 1$ ) marginals:

$$\mathcal{B} = \left\{ \beta_t = \exp \left( -\frac{\zeta^2 + \eta_t^2}{2\alpha} \right) : t \in \mathcal{T} \right\}$$

where  $\zeta \sim N(0, 1)$  and, independently,  $\{\eta_t : t \in \mathcal{T}\}$  arises from a time series model with  $N(0, 1)$  marginals

- Build model for the  $\{\beta_{r,t} : t \in \mathcal{T}\}$  from  $\beta_{r,t} = \exp\{-(\zeta_r^2 + \eta_{r,t}^2)/(2\alpha)\}$ 
  - ▷  $\zeta_r \stackrel{ind.}{\sim} N(0, 1)$
  - ▷ AR(1) process for the  $\{\eta_{r,t} : t \in \mathcal{T}\}$ :  $\eta_{r,t} \mid \eta_{r,t-1}, \phi \sim N(\phi\eta_{r,t-1}, 1 - \phi^2)$  with  $|\phi| < 1$  (and  $\eta_{r,1} \stackrel{ind.}{\sim} N(0, 1)$ )
- Different types of correlations can be studied:  $\text{corr}(p_{r,t}, p_{r,t+k})$ , and  $\text{corr}(G_t(A), G_{t+1}(A))$ , for any subset  $A$  in the support of the  $G_t$

# DDP model extension

- Stochastic process with beta( $\alpha, 1$ ) marginals:

$$\mathcal{B} = \left\{ \beta_t = \exp \left( -\frac{\zeta^2 + \eta_t^2}{2\alpha} \right) : t \in \mathcal{T} \right\}$$

where  $\zeta \sim N(0, 1)$  and, independently,  $\{\eta_t : t \in \mathcal{T}\}$  arises from a time series model with  $N(0, 1)$  marginals

- Build model for the  $\{\beta_{r,t} : t \in \mathcal{T}\}$  from  $\beta_{r,t} = \exp\{-(\zeta_r^2 + \eta_{r,t}^2)/(2\alpha)\}$ 
  - ▷  $\zeta_r \stackrel{ind.}{\sim} N(0, 1)$
  - ▷ AR(1) process for the  $\{\eta_{r,t} : t \in \mathcal{T}\}$ :  $\eta_{r,t} \mid \eta_{r,t-1}, \phi \sim N(\phi\eta_{r,t-1}, 1 - \phi^2)$  with  $|\phi| < 1$  (and  $\eta_{r,1} \stackrel{ind.}{\sim} N(0, 1)$ )
- Different types of correlations can be studied:  $\text{corr}(p_{r,t}, p_{r,t+k})$ , and  $\text{corr}(G_t(A), G_{t+1}(A))$ , for any subset  $A$  in the support of the  $G_t$

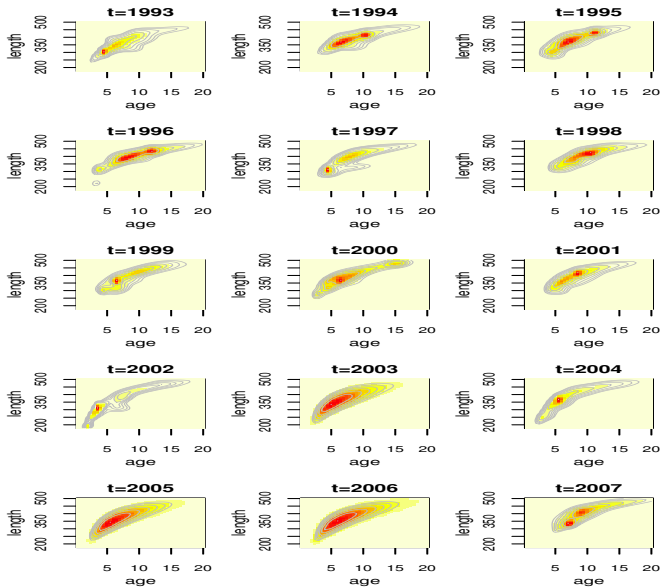
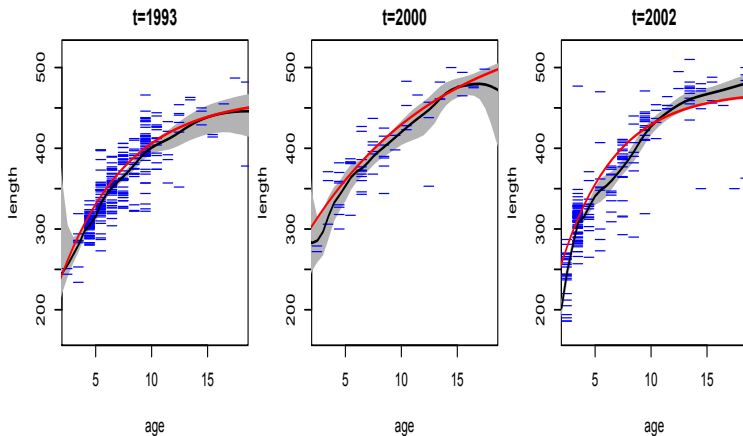
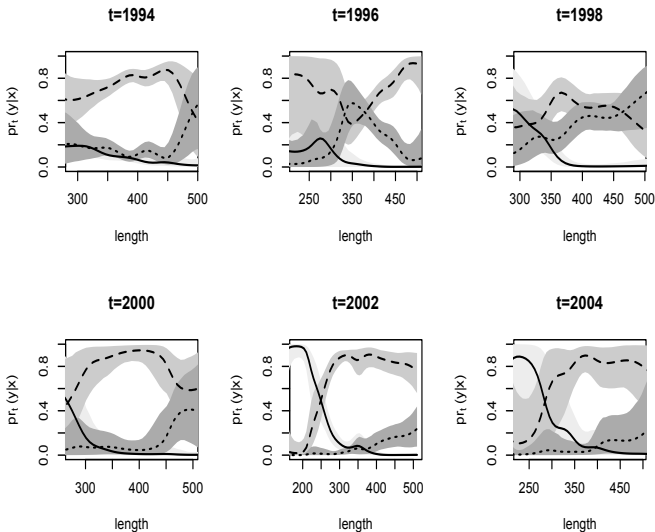


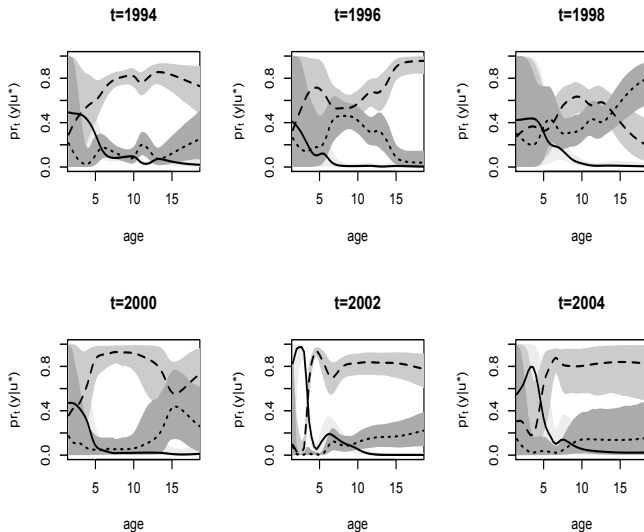
Figure: Posterior mean estimates for  $f(\text{age}, \text{length})$ .



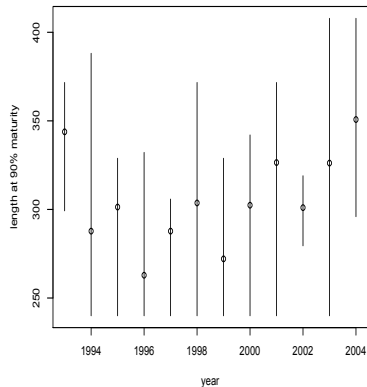
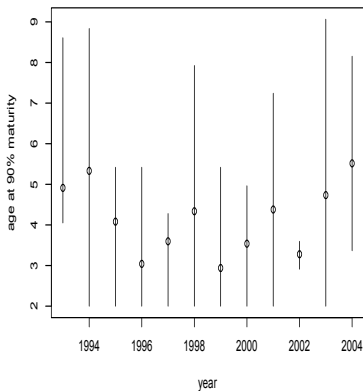
**Figure:** Posterior mean and 95% interval bands for the expected value of length over (continuous) age, across three years. Overlaid are the data (in blue) and the estimated von Bertalanffy growth curves (in red).



**Figure:** Posterior mean and 95% interval bands for the maturation probability curves associated with length: immature (solid); pre-spawning mature (dashed); post-spawning mature (dotted).



**Figure:** Posterior mean and 95% interval bands for the maturation probability curves associated with age: immature (solid); pre-spawning mature (dashed); post-spawning mature (dotted).



**Figure:** Posterior mean and 90% intervals for the smallest value of age above 2 years at which probability of maturity first exceeds 90% (left), and similar inference for length (right).

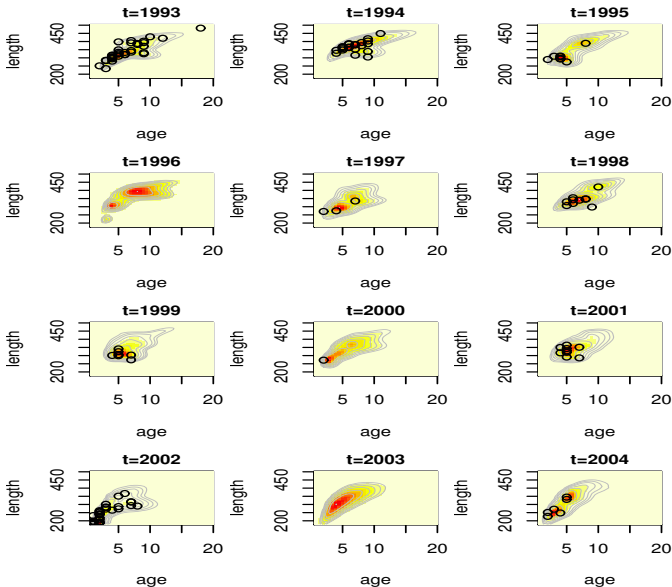
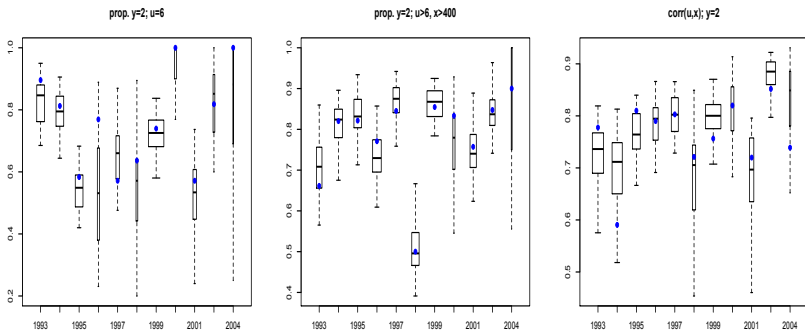


Figure: Posterior mean estimates for  $f(\text{age}, \text{length} \mid Y = 1)$ , with corresponding data overlaid.



**Figure:** Results from posterior predictive model checking. Proportion of age = 6 pre-spawning mature fish (left), proportion of age  $\geq 7$ , and length  $> 400$  mm pre-spawning mature fish (middle), and sample correlation between length and age for pre-spawning mature fish. The blue circles in the left and middle panels denote the actual data proportions, and in the right panel the data-based correlation.

## References/Acknowledgments

- Modeling framework for ordinal regression problems with a small to moderate number of covariates, and for settings where modeling the joint response-covariate distribution is appropriate (or necessary).
  - ▷ DeYoreo, M. & Kottas, A. (2017). "Bayesian nonparametric modeling for multivariate ordinal regression." *JCGS*.
  - ▷ DeYoreo, M. & Kottas, A. (2017). "Modeling for dynamic ordinal regression relationships: An application to estimating maturity of rockfish in California." *JASA*.
- Thanks to: Stephan Munch, Alec MacCall, Don Pearson, Marc Mangel
- Funding from NSF under award DMS 1310438

## References/Acknowledgments

- Modeling framework for ordinal regression problems with a small to moderate number of covariates, and for settings where modeling the joint response-covariate distribution is appropriate (or necessary).
  - ▷ DeYoreo, M. & Kottas, A. (2017). "Bayesian nonparametric modeling for multivariate ordinal regression." *JCGS*.
  - ▷ DeYoreo, M. & Kottas, A. (2017). "Modeling for dynamic ordinal regression relationships: An application to estimating maturity of rockfish in California." *JASA*.
- Thanks to: Stephan Munch, Alec MacCall, Don Pearson, Marc Mangel
- Funding from NSF under award DMS 1310438

## References/Acknowledgments

- Modeling framework for ordinal regression problems with a small to moderate number of covariates, and for settings where modeling the joint response-covariate distribution is appropriate (or necessary).
  - ▷ DeYoreo, M. & Kottas, A. (2017). "Bayesian nonparametric modeling for multivariate ordinal regression." *JCGS*.
  - ▷ DeYoreo, M. & Kottas, A. (2017). "Modeling for dynamic ordinal regression relationships: An application to estimating maturity of rockfish in California." *JASA*.
- Thanks to: Stephan Munch, Alec MacCall, Don Pearson, Marc Mangel
- Funding from NSF under award DMS 1310438

MANY THANKS !