Australia-Japan Workshop on Data Science 2009
Keio University, Yagami campus
March 24 – 27

Statistical challenges for modeling data with many zeros

Mihoko Minami
The Institute of Statistical Mathematics
Keio University (from April, 2009)

Joint research with
Dr. Cleridy E. Lennert-Cody
Inter-American Tropical Tuna Commission

Statistical challenges for modeling data with many zeros

- Count data with many zero-valued observations
  - The number of defects in a manufacturing process (Lambert, 1992)
  - The number of days per time period that people of working age missed their primary activities due to illness (Lam et al. 2006)
  - The number of animals per unit area or unit of effort (Welsh et al.1996)
- Analyzing such data without any consideration for excess zeros may produce misleading results
- Menu
  - Possible consequence of ignoring excess zeros in analysis
  - A new feature extraction method for very non-normal multivariate data, such as multivariate data with many zero-valued observations

Overestimation of trend by the negative binomial regression model fitted to data with many zeros

- A negative binomial regression model is often used to analyze count data with over-dispersion relative to Poisson model.
  - The negative binomial regression model with log link with size parameter θ (cf. Lawless, 1987)

$$f_{NB}(y;\mu,\theta) = \frac{\Gamma(\theta+y)}{\Gamma(\theta)\Gamma(y+1)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{y} \quad \text{where} \quad \log(\mu) = \mathbf{x}^T\boldsymbol{\beta}$$

  - mean and variance $\mathrm{E}[Y \mid \mu,\theta] = \mu, \quad \mathrm{Var}[Y \mid \mu,\theta] = \mu + \frac{1}{\theta}\mu^2$
- The negative binomial distribution is flexible
- It may appear adequate for datasets with many zero valued observations.

However, the negative binomial regression model may produce very misleading results for data with excess zeros.

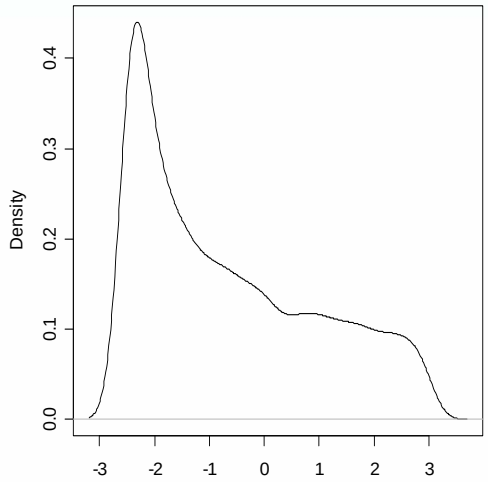
Overestimation of trend by negative binomial regression model

Simulation setting :

Generate negative binomial data *w*

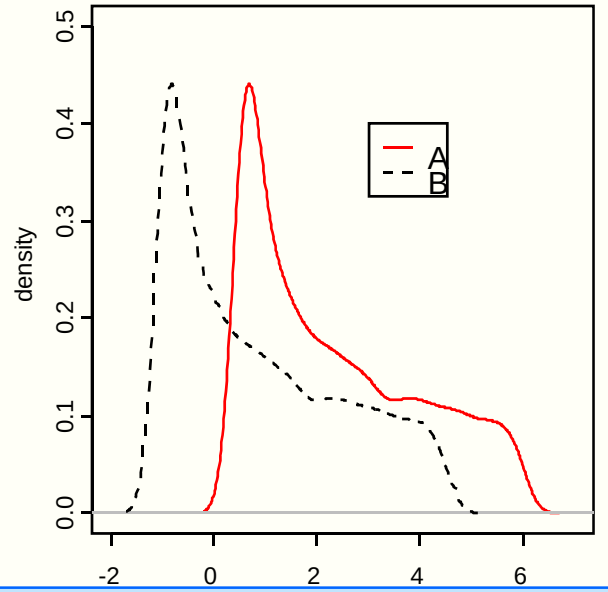
- generate $w_i \sim$ negative binomial regression model
  with a factor of two levels A and B and covariate x_i
  with size parameter $\theta = 0.6.$

density of x_i



$$\log(\mu_i) = x_i + 3 - 1.5\, I_B(i)$$

density of $\log(\mu_i)$



Sample sizes are 10,000 for level A and B (20,000 in total)

Overestimation of trend by negative binomial regression model

Simulation setting :

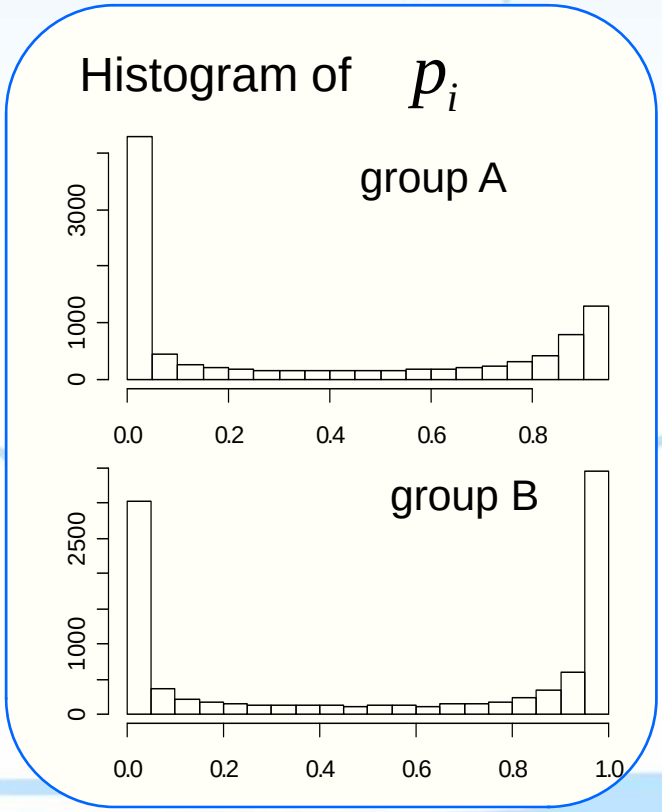
Replace data with zeros to get thinned data **y**

- Random values y_i were obtained by replacing w_i by value zero with probability p_i where

$$\log\left(\frac{p_i}{1-p_i}\right) = -3x_i - 5 + 2.5\, I_B(i)$$

- The numbers of zeros (out of 10,000)

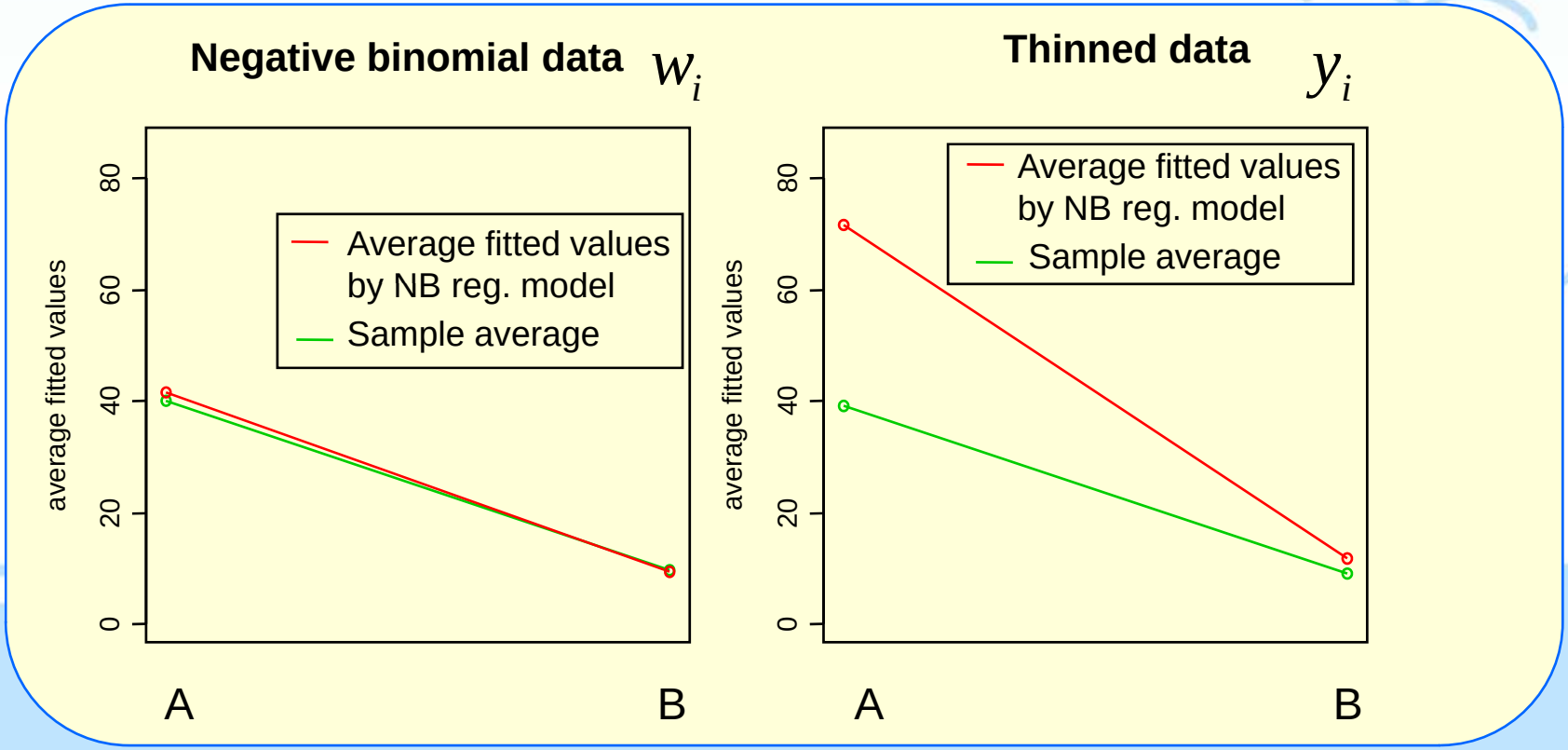
Level	Neg. Bin. data w	Thinned data y	Replaced by zeros
A	2333	4531	1867
B	4466	6333	2198



Overestimation of trend by negative binomial regression

Average fitted values

- Negative binomial regression model was fitted to both data.
- For level A, Average of $\hat{w}_i$ << Average of $\hat{y}_i$ even though $w_i \geq y_i$ by its construction



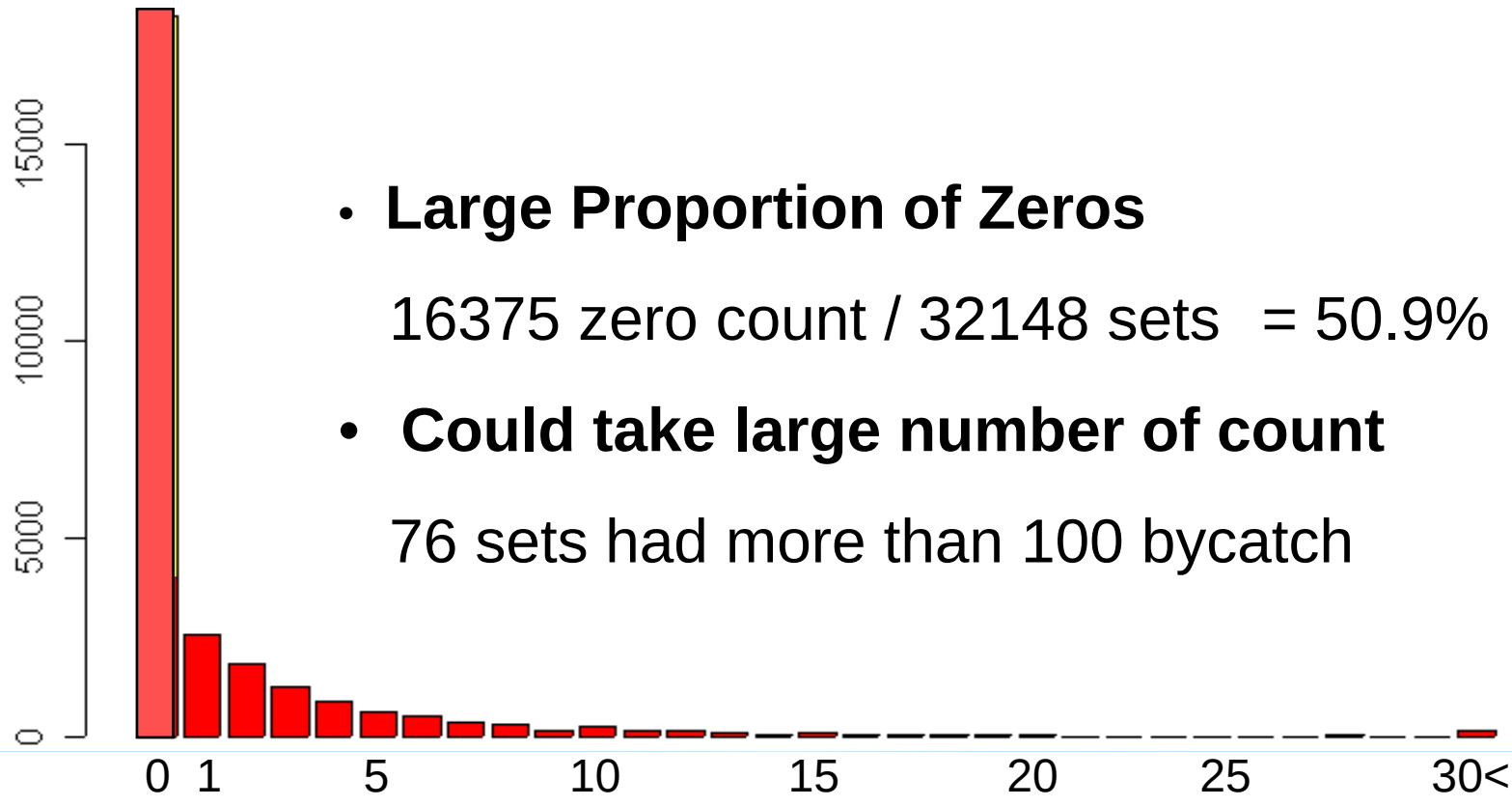
Shark bycatch data from a tuna purse-seine fishery

Purse-seine fishery

Floating object

- Data on floating object sets between 1994 – 2004 collected by IATTC observers onboard large tuna vessels of the international purse-seine fleet
- Sample size 32,148
- Variables recorded for each set
  - **Silky shark bycatch count** (response)
  - **Environmental conditions**: Location, year, calendar date, sea surface temperature
  - **Operational conditions**: Net depth, floating object depth, amount of tuna（log), Amount of non-silky shark bycatch (log),Measurement on floating objects such as number of unique object numbers within 5 degree square

Histogram of silky shark bycatch count per set between 1994 and 2004



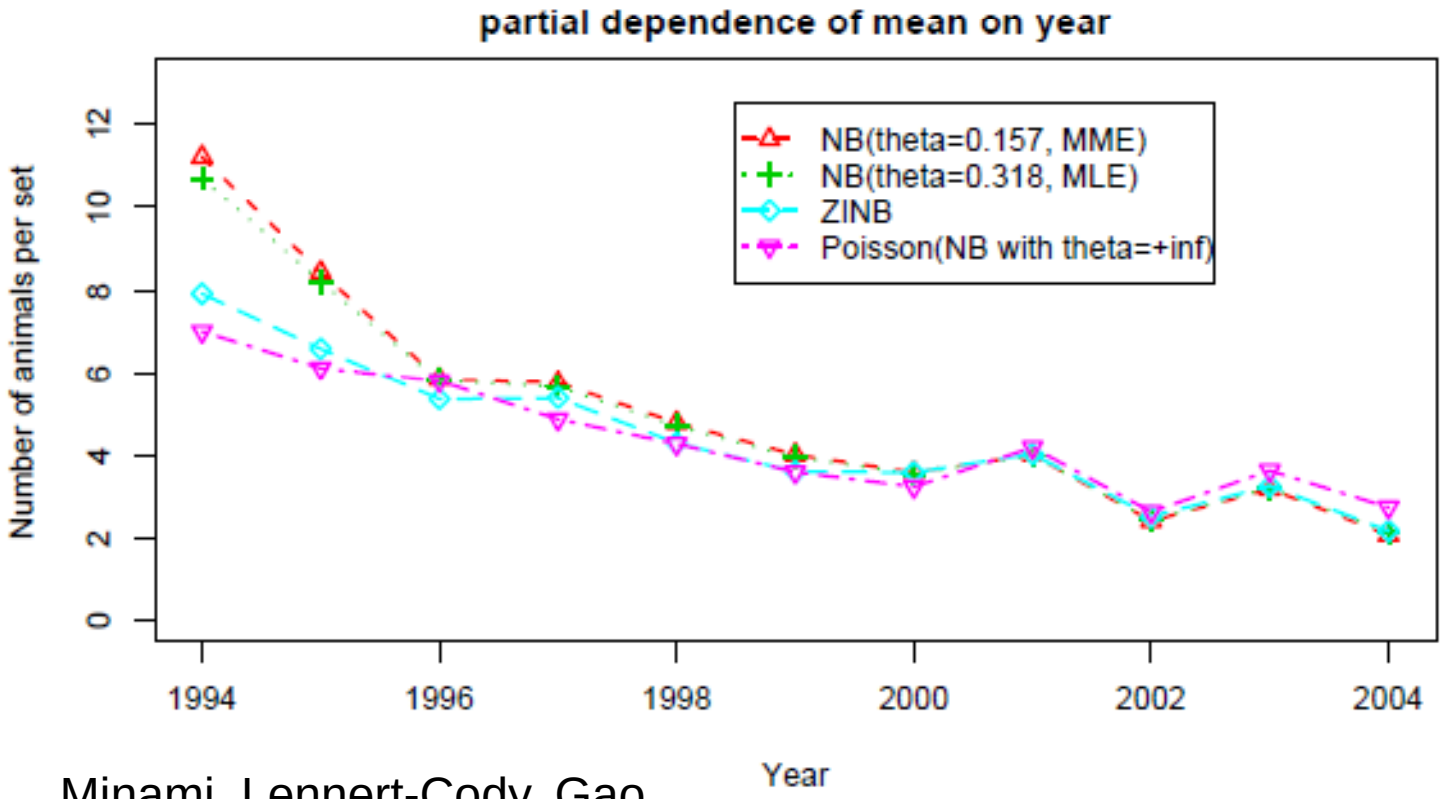
- **Large Proportion of Zeros**

  16375 zero count / 32148 sets = 50.9%

- **Could take large number of count**

  76 sets had more than 100 bycatch

Trend in the standardized silky shark bycatch
(Partial dependence plot, Hastie et al 2001)



Minami, Lennert-Cody, Gao and Román-Verdesoto (2007)

Partial dependence

$$f_s(X_s) = \frac{1}{n}\sum_{i=1}^{n} f(X_s, x_{iC})$$

where

$\{x_{1C}, x_{2C}, \cdots, x_{nC}\}$

are values of other covariates occurring In the dataset

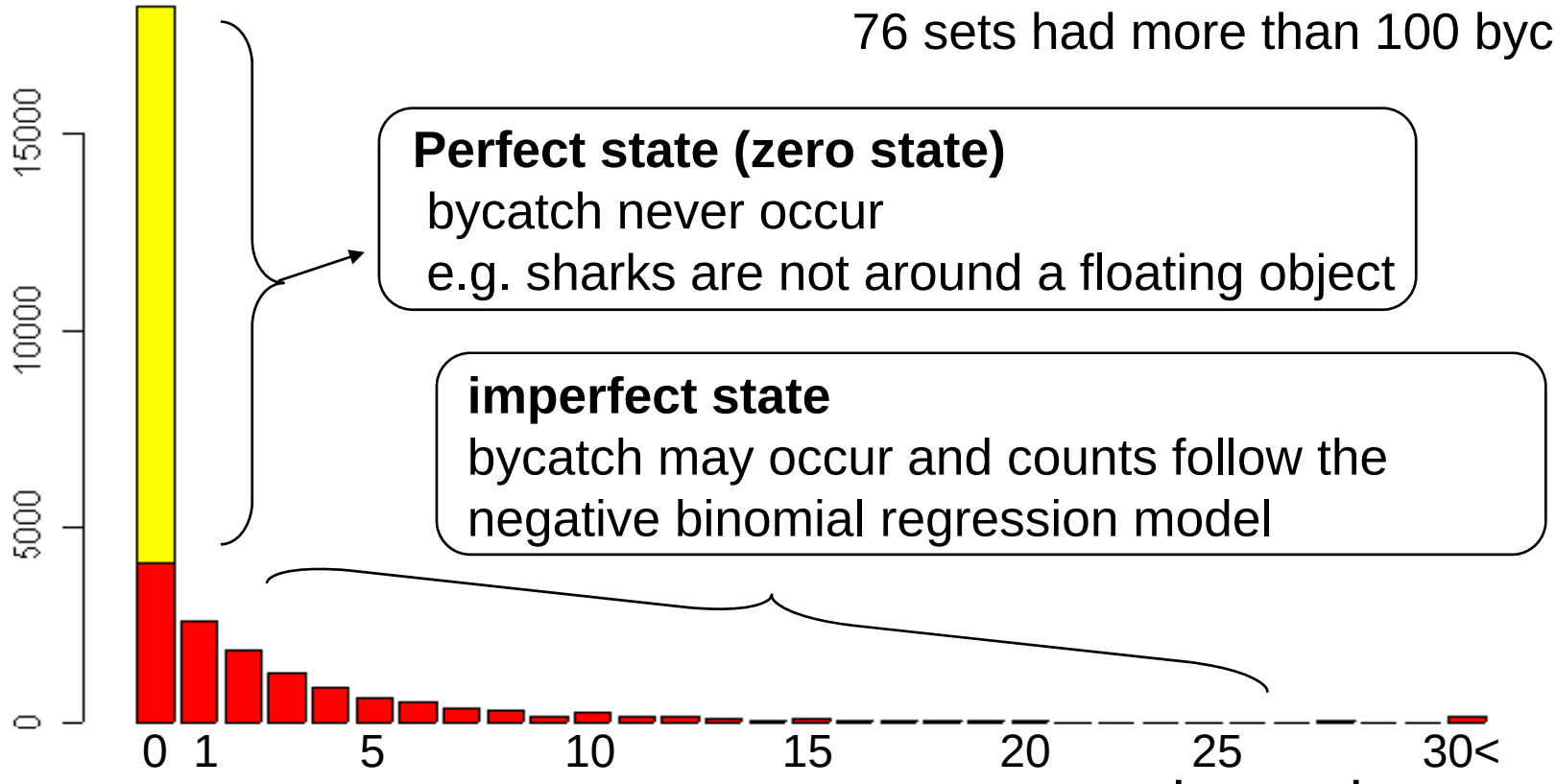
Sample size: 32148

Negative binomial regression model overestimated temporal trend

Modeling shark bycatch: The Zero-Inflated Negative Binomial (ZINB) Regression Model (with Smoothing)

Histogram of silky shark bycatch counts

- **Large Proportion of Zeros**
  16375 zero count / 32148 sets = 50.9%
- **Could take large number of count**
  76 sets had more than 100 bycatch



Perfect state (zero state)
bycatch never occur
e.g. sharks are not around a floating object

imperfect state
bycatch may occur and counts follow the negative binomial regression model

Overestimation of trend by negative binomial regression model

Models for simulated data

- Thinned data Y_i follow zero-inflated negative binomial regression model ZINB(μ_i, p_i, θ)
  - probability function

$$f(y;\mu_i, p_i, \theta) = \begin{cases} p_i + (1-p_i) f_{NB}(y=0;\mu_i,\theta) & y=0 \\ (1-p_i) f_{NB}(y;\mu_i,\theta) & y=1,2,\ldots \end{cases}$$

where

$$f_{NB}(y;\mu,\theta) = \frac{\Gamma(\theta+y)}{\Gamma(\theta)\Gamma(y+1)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{y} \quad \theta>0, \lambda>0$$

$$\log\left(\frac{p_i}{1-p_i}\right) = -3x_i - 5 + 2.5I_B(i) \quad \text{and} \quad \log(\mu_i) = x_i + 3 - 1.5I_B(i)$$

Overestimation of trend by negative binomial regression model

When does overestimation occur?

1. **The data contain many zero-valued observations**, and can be considered to follow a ZINB regression model.
2. **The probability of an observation being in the zero state (perfect state) depends on covariates**.
3. **Positive counts are over dispersed** relative to a Poisson model.
4. **The variable that increases the probability of being in the zero-state decreases the mean of the incomplete state.**
   That is, the signs of coefficients in the negative binomial regression part and the logistic regression part of a ZINB regression model are opposite of each other .

Overestimation of trend by negative binomial regression model

How does overestimation occur?

1. **The estimate of the size parameter becomes much smaller** than the value of the size parameter in the underlying ZINB regression model

2. **The estimated coefficients becomes larger in absolute value** than the corresponding coefficients in the negative binomial regression part of the underlying ZINB regression model.

3. **The fitted values are quite large** when the covariates have large absolute values and their signs are the same as the coefficients.

Overestimation of trend by negative binomial regression model

Why does overestimation occur?

1. The estimates of the size parameter becomes quite small

Property 1 *When* $NB(\mu, \theta)$ *model is fitted to samples drawn from* $ZINB(p, \mu_0, \theta_0)$ *with* $0 < p < 1$, *as the sample sizes increase to* $+\infty$

1. *The method of moments estimate* $\hat{\theta}_{MM}$ *converges to* θ^* *where*

$$\theta^* = \theta_0 \left(\frac{1-p}{1+\theta_0 p} \right) \quad \text{and} \quad \theta^* < \theta_0.$$

2. *The maximum likelihood estimate* $\hat{\theta}_{MLE}$ *converges to* $\theta^\dagger$ *where* $\theta^\dagger < \theta_0$.

For simulated data, $\hat{\theta}_{\text{MLE}} = 0.42$ and $\hat{\theta}_{\text{MME}} = 0.305$ while $\theta = 0.6$.

For shark bycatch data, $\hat{\theta}_{\text{MLE}} = 0.318$ and $\hat{\theta}_{\text{MME}} = 0.157$ while $\hat{\theta}_{ZINB} = 0.568$

Overestimation of trend by negative binomial regression model

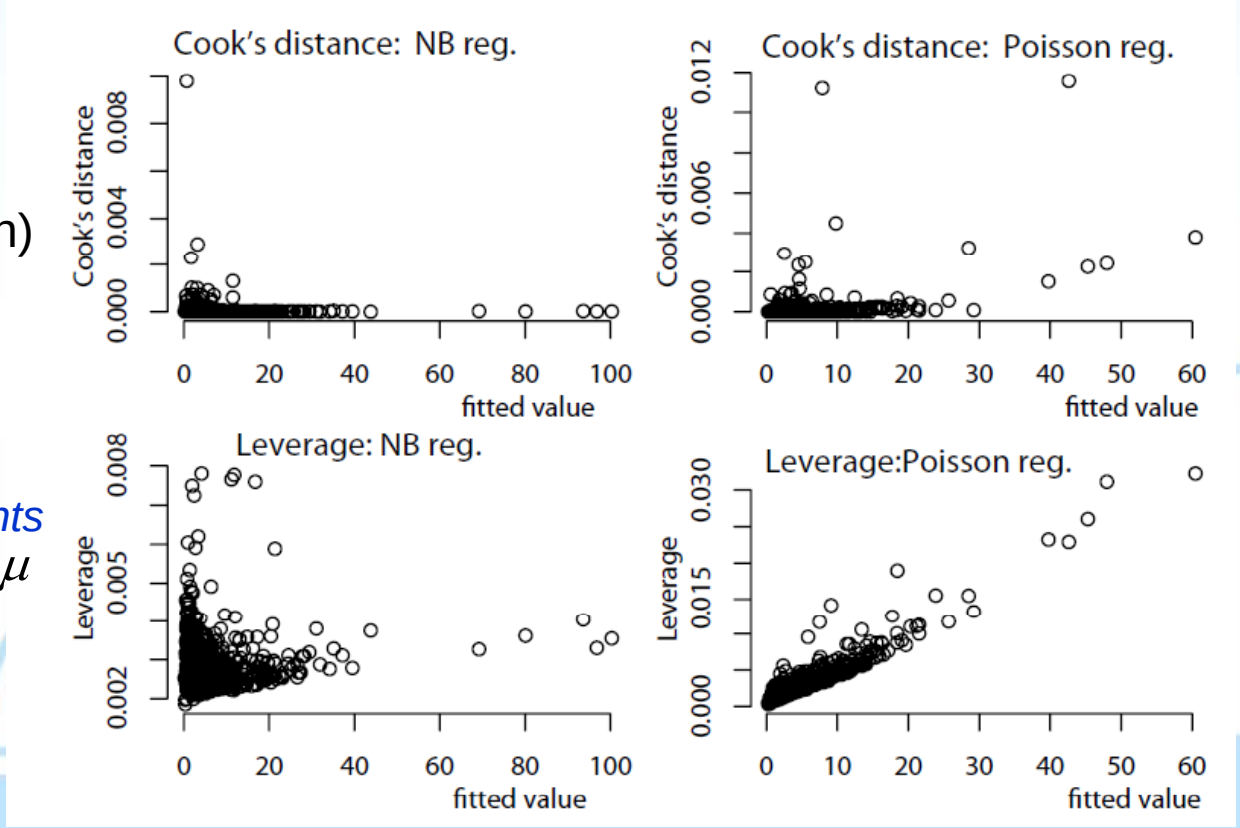
Why does overestimation occur?

2. ML estimator puts more weight on observations with small fitted values

The Score function :
(The influence function is proportional to the score function)

$$\frac{\partial l_{\mathrm{NB}}(\beta,\theta|y_i,B_i)}{\partial\beta} = (y_i-\mu_i)\cdot\left(\frac{\theta}{\theta+\mu_i}\right)B_i$$

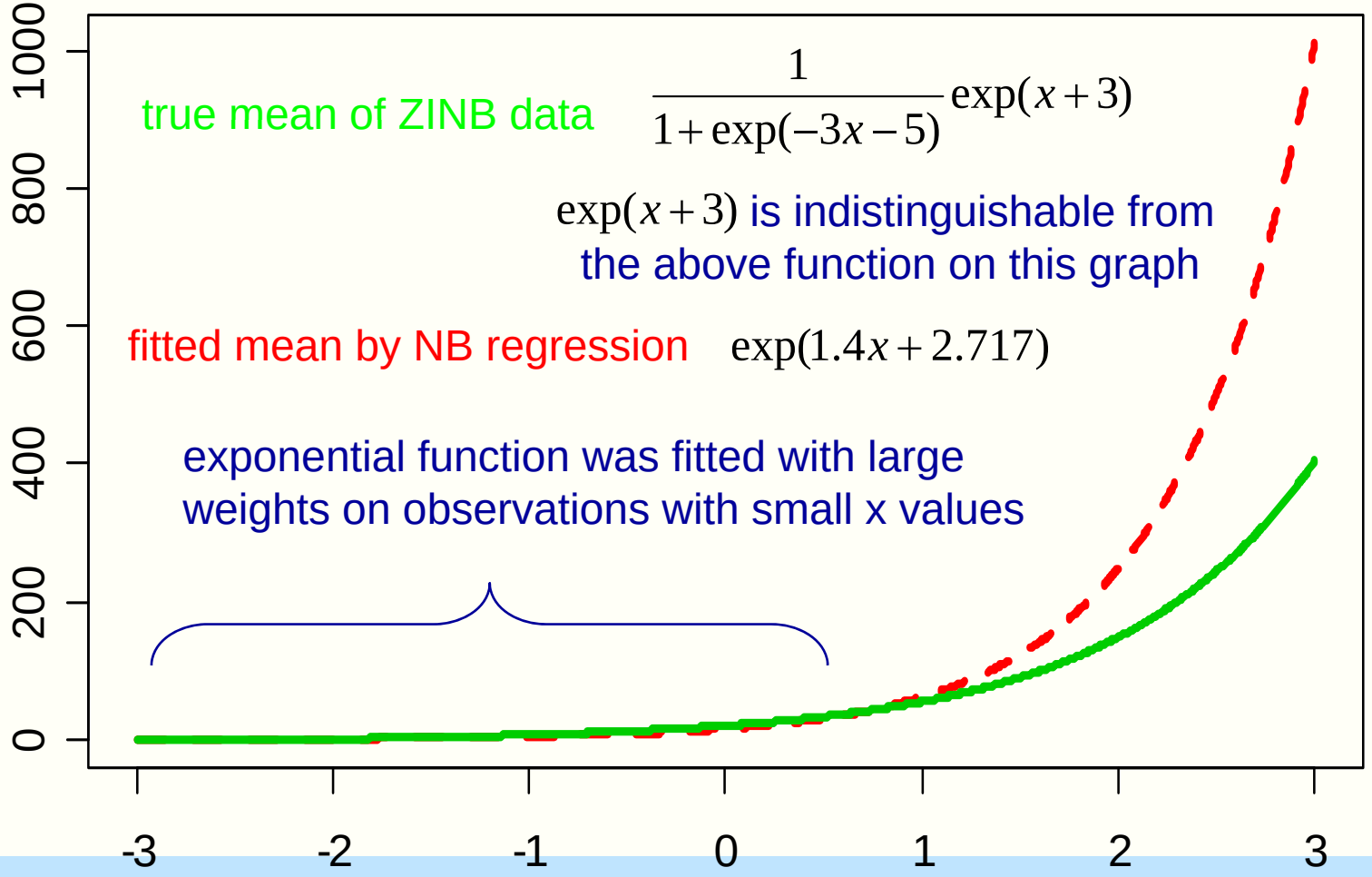
For small θ, relative weights decrease more rapidly as μ increases



Overestimation of trend by negative binomial regression model

Why does overestimation occur?

3. Localized fitting causes overestimation of coefficients

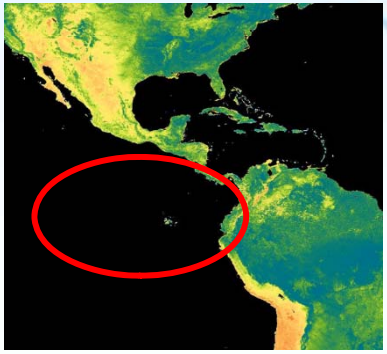


A New Feature Extraction Method for Very Non-Normal Data

- We propose a new feature extraction method for very non-normal data such as count data with many zeros.
- Our method extends principal component analysis (PCA) in the same manner as the generalized linear model (GLM) extends the ordinary linear regression model (LM).
- We apply this method to multivariate species-size data from purse-seine tuna fisheries in eastern Pacific Ocean.
- These data contain many zero-valued observations for each variable (combinations of species and size).
- Thus, as an error distribution we use Tweedie distribution which has a probability mass at zero and apply Tweedie-GPCA method to the data.

Multivariate species-size Data in Purse-seine Tuna Fisheries

- Data were collected by IATTC (Inter-American Tropical Tuna Commission) observers onboard large tuna vessels of the international purse-seine fleet in the eastern Pacific Ocean in 2000.



Purse-seine fisheries

- For each floating-object set, the amount/count of catch and bycatch are recorded by species and size interval.
  There are 56 variables × 2834 sets.

Using this dataset we would like to explore species associations and possible relationships between these associations and the environmental and fishery operational factors.

Catches & bycatches (56 species-sizes)

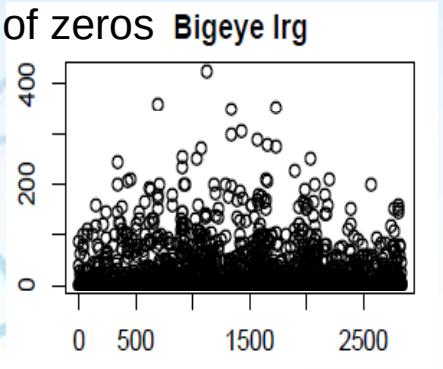
- Target species

Skipjack

Yellowfin tuna

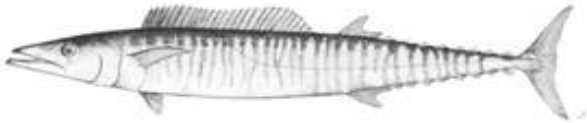
BIgeye tuna

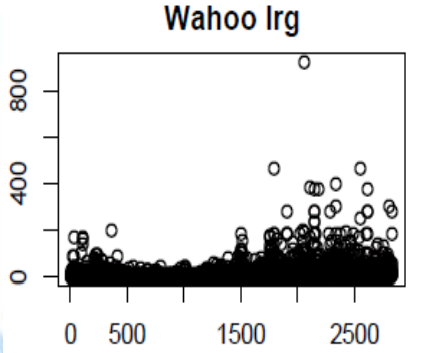
55% of zeros



- Bycatch species targeted in other fisheries

Wahoo

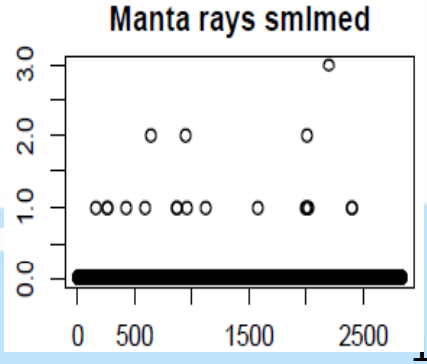




34% of zeros

- Potentially vulnerable bycatch species targeted in other fisheries
- Vulnerable bycatch species

Manta ray



99.5% of zeros

- Other bycatch species

Feature extraction methods

Statistical model

$$= \left[\text{Systematic effects} \;,\; \text{Random effects} \right]$$

Focus is mainly on

Existing methods:
principal component analysis(PCA),
independent component analysis (ICA)
non-negative matrix factorization(NMF),
principal curves, non-linear PCA,
kernel PCA

We want to pay more attention to the random effect part.

Species-size data cannot be transformed to nearly Gaussian with any transformation

metric/non-metric multidimensional scaling methods

PCA (principal component analysis)

For simplicity, here we assume the mean of each variable is zero or the sample average is subtracted from the observations.

- Let $\mathbf{Y}$ be data matrix of N samples and m variables, and $\mathbf{a}_i, i = 1, \cdots, m$ denote coefficient vectors for the i th principal component.
- Among matrices of size $m \times k$ with orthonormal columns, matrix $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k)$ maximizes the total variance, i.e.

$$\mathrm{tr}(\mathrm{Var}(\mathbf{YA})) = \max_{\mathbf{B}^{\mathrm{T}}\mathbf{B}=\mathbf{I}_k} (\mathrm{tr}(\mathrm{Var}(\mathbf{YB}))$$

- This is equivalent that $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k)$ minimizes the mean square reconstruction error, i.e. $\sum_{i=1}^{N} \sum_{j=1}^{m} E_{ij}^2$ where $\mathbf{E} = \mathbf{Y} - \mathbf{YBB}^{\mathrm{T}}$ and $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}_k$ is minimized when $\mathbf{B} = \mathbf{A}$

PCA (principal component analysis)

- Suppose we extract the first k principal component
- If we express $\mathbf{X} = \mathbf{YA}$ then $\mathbf{YAA}^{\mathbf{T}} = \mathbf{XA}^{\mathbf{T}}$.
- PCA maximizes the likelihood under the model

$$\mathbf{Y} = \mathbf{X}\mathbf{A}^{T} + \mathbf{E}, \qquad \mathbf{E}_{ij} \sim N(0,\ \sigma^{2}),\ i.i.d.$$

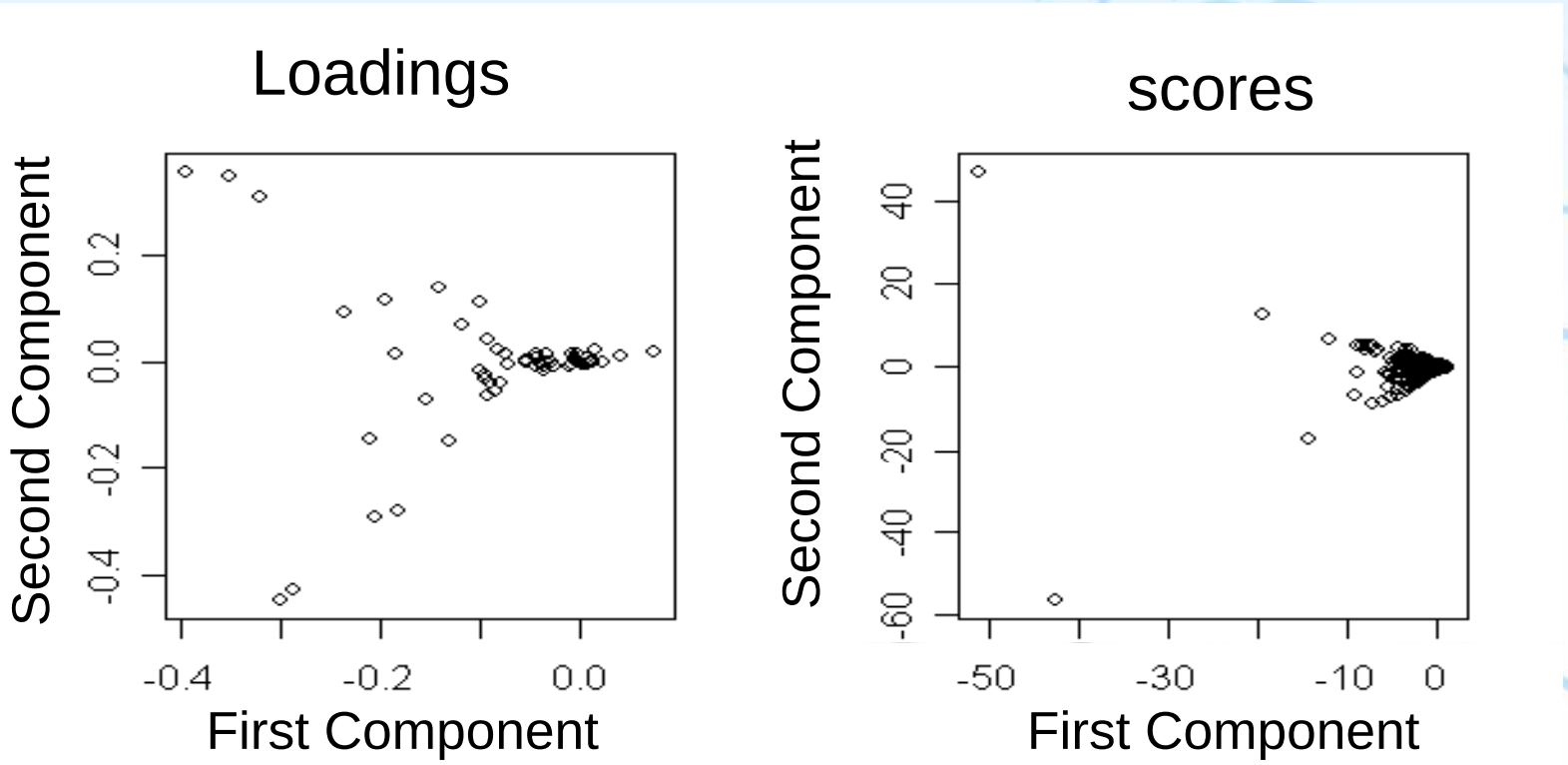
where $\mathbf{X}$ and $\mathbf{A}$ are matrices of size $N \times k$ and $m \times k$

or equivalently, under the model

$$\mathbf{Y} = \mathbf{M} + \mathbf{E}, \qquad \mathbf{E}_{ij} \sim N(0,\ \sigma^{2}),\ i.i.d.$$

where $\mathbf{M}$ is a matrix of rank k .

If we apply PCA to species-size data



A few points have strong influence on results. Even if these points are omitted, the plots look the same. The percent of the variance explained by each of the first two components is rather small: **5.1%** and **4.6%.**

Generalized PCA (GPCA)

We propose a **generalized PCA** (GPCA) method that extends principal component analysis in the following sense:

1. The rank of matrix $g(\mathbf{M})$, rather than mean matrix $\mathbf{M}\ (=\mathrm{E}(\mathbf{Y}))$ itself, is k where g is a monotone increasing function.
2. Y_{ij} independently follows a distribution $f(y;\mathbf{M}_{ij},\sigma^2)$ in exponential family.

GPCA extends PCA in the same manner as the generalized linear model (GLM) extends the linear regression model (LM):

1. $g(\mathbf{m})$ rather than $\mathbf{m}\ (=\mathrm{E}(\mathbf{Y}))$ is a linear function of covariates
2. Y_i independently follows a distribution $f(y;\mathbf{m}_i,\sigma^2)$ in exponential family.

Generalized PCA (GPCA)

- We first find the matrix $\mathbf{G}$ of rank k which is the transformed mean matrix $g(\mathbf{M})$, for $\mathbf{M}[=\mathrm{E}(\mathbf{Y})]$ and link function g .
- Then, we find the basis of $\mathbf{G}$ that shows characteristic features of individuals (sets in species-size data) and variables.
- To find them, we use singular value decomposition (SVD) or Independent component analysis (ICA) .

GPCA:
Estimation algorithm for matrix G (=g(M))

- We decompose as $\mathrm{G} = \mathbf{X}\mathbf{A}^T$ where $\mathbf{X} \in \Re^{N \times k}$ and $\mathbf{A} \in \Re^{m \times k}$ and maximize the likelihood by alternately fixing $\mathbf{X}$ or $\mathbf{A}$ and updating the other.
- Maximizing the likelihood with fixed $\mathbf{X}$ or $\mathbf{A}$ is reduced to find the MLEs for m or N generalized linear models.
  - When $\mathbf{X}$ is given, for $j = 1, \cdots, m$,

$$g(\mathbf{m}_j) = \mathbf{X} A_j^T$$

    where $\mathbf{m}_j = \mathrm{E}(\mathbf{Y}_j)$, $\mathbf{Y}_j$ is the j th column of $\mathbf{Y}$ and A_j is the j th row of $\mathbf{A}$
  - When $\mathbf{A}$ is given, for $i = 1, \cdots, N$,

$$g(\tilde{\mathbf{m}}_i^T) = \mathbf{A} X_i^T$$

    where $\tilde{\mathbf{m}}_i = \mathrm{E}(\tilde{\mathbf{Y}}_i)$, $\tilde{\mathbf{Y}}_i$ is the i th row of $\mathbf{Y}$ and X_i is the i th row of $\mathbf{X}$

GPCA:
Estimation algorithm for matrix G (=g(M))

Algorithm : Repeat the following steps:

.

1. $\mathbf{X}$ is fixed; $g(\mathbf{m}_j) = \mathbf{X} A_j^T$ For $j = 1, \cdots, m$, update A_j using one IRLS (iterative re-weighted least square) update under the generalized linear model with response vector $\mathbf{Y}_j$, covariate matrix $\mathbf{X}$ and link function g.

2. $\mathbf{A}$ is fixed; $g(\tilde{\mathbf{m}}_i^T) = \mathbf{A} X_i^T$ For $i = 1, \cdots, N$, update X_i using one IRLS update under the generalized linear model with response vector $\tilde{\mathbf{Y}}_i$, covariate matrix $\mathbf{A}$ and link function g.

GPCA: Characteristic features (loadings and scores)

Let $\mathrm{G} = \mathbf{X}\mathbf{A}^T$. G is a matrix of size $\mathrm{N} \times \mathrm{m}$ and with rank k.

- By singular value decomposition (SVD):

$$\mathrm{G} = \mathbf{U}\Sigma\mathbf{V}^T \text{ where } \mathbf{U} \in \Re^{N\times k}, \mathbf{V} \in \Re^{m\times k}, \mathbf{U}^T\mathbf{U} = \mathbf{I}_k, \mathbf{V}^T\mathbf{V} = \mathbf{I}_k$$

$$\text{and} \quad \Sigma = \mathrm{diagonal}(\lambda_1, \cdots, \lambda_k), \quad \lambda_1 \geq \cdots \geq \lambda_k$$

  we consider the columns of $\mathbf{U}$ as scores for individuals (sets in species-size data) and the columns of $\mathbf{V}$ as loadings for variables.

- By independent component analysis(ICA):
  Perform ICA with $\mathbf{U}$ as observation matrix.
  We consider obtained independent components (the columns of $\mathbf{S} = \mathbf{U}\mathbf{W}$ where $\mathbf{W}$ is the recovering matrix) as scores for individuals and the columns of $\mathbf{T} = \mathbf{V}\Sigma\mathbf{W}^{-T}$ as loadings for variables.

Tweedie distribution

- Tweedie distribution $\mathrm{Tw}_p(\mu, \sigma^2)$ with $1 < p < 2$ is a compound Poisson distribution and belongs to exponential family. It is the distribution of the following Y :

$$Y = \sum_{i=1}^{Z} X_i \quad \text{where} \quad Z \sim \text{Poisson}(m),$$
$$X_i \sim \text{Gamma}(m_g, \sigma_g^2) \quad (\text{mean } m_g, \text{variance } \sigma_g^2 m_g^2)$$

- Tweedie distribution has a probability mass at 0 and the density function is given by

$$p(y;\mu,\sigma^2,p) = a(y;\sigma^2,p)\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu,p)\right\}$$

where
$$d(y;\mu,p) = 2\left\{\frac{y^{2-p}}{(1-p)(2-p)} - \frac{y\mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p}\right\}$$

- The variance is $\mathrm{Var}(Y) = \sigma^2\mu^p$

Tweedie-GPCA

For species-size data, we consider Tweedie-GPCA method that uses Tweedie distribution for error and log link function.

$$Y_{ij} \sim \mathrm{Tw}_p(M_{ij}, \sigma^2), \quad \mathrm{rank}(\log(\mathbf{M})) = k$$

The objective function to minimize is

$$f(\mathbf{Y}; \mathbf{X}\mathbf{A}^T, p) = \sum_{i=1}^{N}\sum_{j=1}^{m} d(y_{ij}; (\mathbf{X}\mathbf{A}^T)_{ij}, p)$$

where

$$d(y; \mu, p) = 2\left\{\frac{y^{2-p}}{(1-p)(2-p)} - \frac{y\mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p}\right\}$$

Tweedie-GPCA

- Mean structure: Instead of subtracting the sample mean as in PCA, we include a constant β_j for each variable into a model.

$$G\ (= g(\mathbf{M})) = \mathbf{1}_N \boldsymbol{\beta}^T + \mathbf{X}\mathbf{A}^T \text{ where } \boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)$$

Each column of $\mathbf{X}$ is constrained to have mean 0, i.e. $\mathbf{1}_N^T \mathbf{X} = \mathbf{0}$.

$\beta_j\ (j = 1, \cdots, m)$ are also estimated simultaneously.

The model can be expressed as

$$G = \tilde{\mathbf{X}}\tilde{\mathbf{A}}^T \text{ where } \tilde{\mathbf{X}} = (\mathbf{1}_N, \mathbf{X}) \text{ and } \tilde{\mathbf{A}} = (\boldsymbol{\beta}, \breve{\mathbf{A}})$$

Tweedie-GPCA

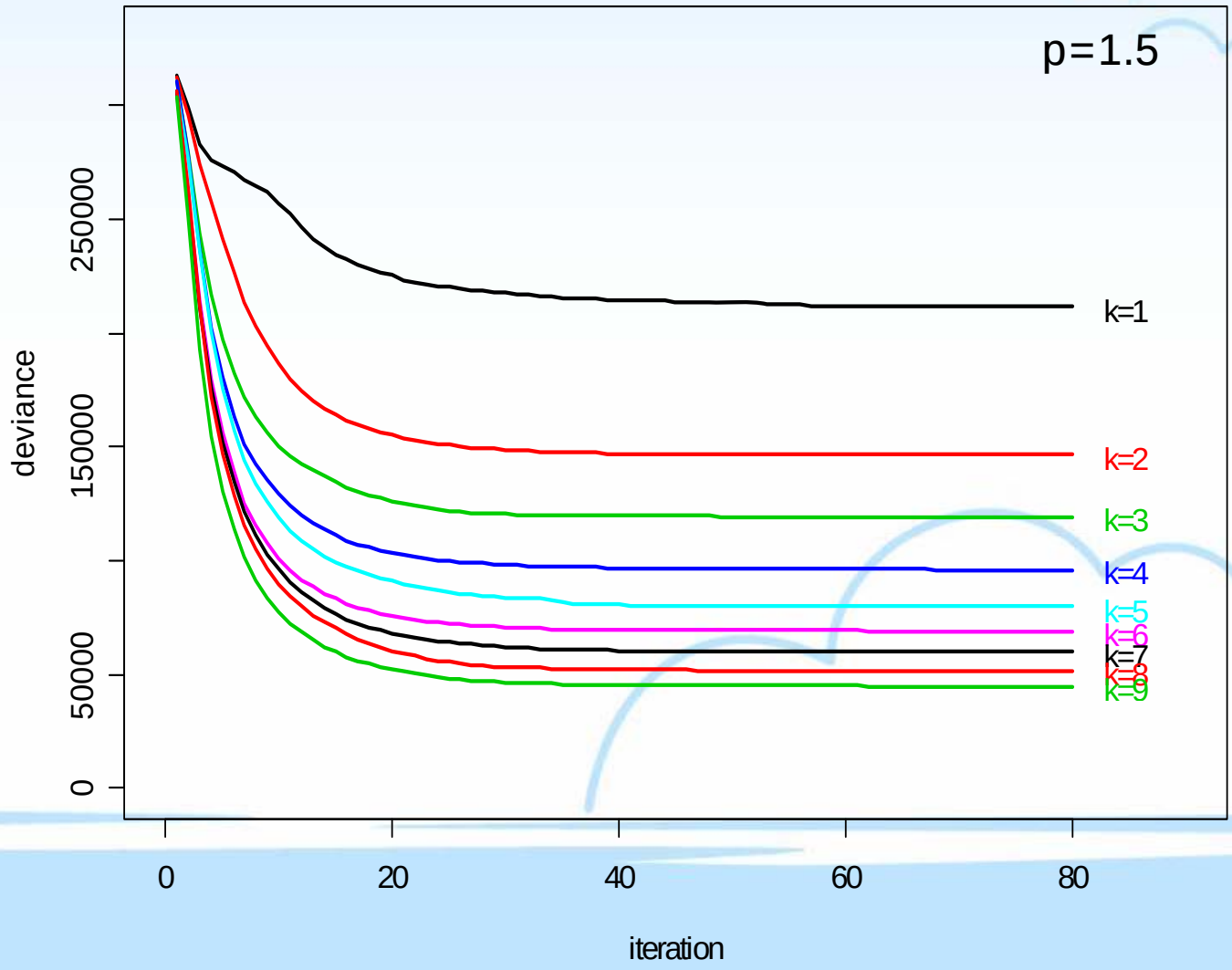
- Scaling: Before estimation, we standardize variables so that deviance estimates for dispersion parameters are equal.

- Proportion of deviance explained : To evaluate the model, we compute the proportion of deviance explained defined:

$$\frac{\text{deviance (null model)} - \text{deviance (candidate model)}}{\text{deviance (null model)}}$$

Tweedie-GPCA:

Sum of deviance versus iteration



k	cumulative % deviance explained	% deviance explained
1	33.3	33.3
2	53.8	20.5
3	62.3	8.5
4	69.5	7.2
5	74.9	5.4
6	78.2	3.3
7	81.1	2.9
8	83.7	2.6
9	85.8	2.1

N=2834, m=56

Features for variables (loadings)

- The loadings of the first feature are all positive. They might be related to overall animal abundance. Target species have very small loadings, which indicates that the catch of target species does not fluctuate much. In all features, species with large percentages of zero catch tend to have loadings with large absolute values.

- The second feature does not show a clear pattern, but filter feeders tend to have large positive values and species that feed on smaller fishes have negative values.

Features for variables (loadings)

- As for the third features, bigeye tuna has the value with large absolute values and there is a possibility this feature is related to amount of catches of bigeye tuna.

- The fourth feature may also relate to target species. Among target species, the loadings of large animal categories are the greatest. Thus, the fourth feature may be related to the size of the target species caught.

Features for sets (scores)

- To see spatial patterns, we averaged features for sets, that is scores, by 1 degree square area. For comparison, we also show the same map for principle component analysis and non-metric multidimensional scaling with Sorensen distance, which is an approach used by ecologists to study association.

- Spatial patterns of the first four features for PCA have similarities, and that means PCA may fail to capture characteristics of the data.

Features for sets (scores)

- It is interesting that the spatial patterns of the features from GPCA and non-metric multidimensional scaling show similarities.However, the spatial pattern of the third feature from GPCA is not found among those from non-metric multidimensional scaling.

- The first two features for variables (species, size) obtained from this method appear to be associated with abundance of several seldom-caught species that are considered vulnerable to fisheries impacts.

Smoothed features for sets (scores)

- The middle row shows smoothed spatial patterns. The first feature shows a hot spot where amounts of bycatch tend to be high. The first feature also shows an overall north-south gradient. The second feature also shows a hot spot. Information on hotspots can be useful to fisheries managers because hotspots may indicate candidate areas for fishery closures.
- The third feature shows a strong in-shore off-shore pattern and the fourth feature may reflect influence of the Peru current. We believe that the spatial patterns of features1, 3 and 4 all show larger-scale structure which may be related to the oceanography of the region.

Conclusion

- We proposed a new feature extraction method, generalized PCA (GPCA), and applied it to multivariate species and size composition data from a tuna purse-seine fishery.
- The results suggest that GPCA may be useful tool for identifying areas within the region occupied by the purse-seine fishery with greater occurrence of bycatch of vulnerable species.
- These results also suggest that, more generally, the Tweedie-GPCA method shows promise as a tool for studying the impact of fisheries on ecosystem and exploring ecosystem community structure

Acknowledgments

We would like to thank the Inter-American Tropical Tuna Commission for providing the data.