# Growth stanzas from kernel mixture models of snow crab size frequency

Jae S. Choi[1,*]

[1]Bedford Institute of Oceanography, Fisheriess and Oceans Canada
[*]jae.choi@dfo-mpo.gc.ca

January 8, 2025

**Abstract**

*Snow crab (*Chionoecetes opilio*) are cold-water stenotherms in the northern hemisphere. As they are long-lived and have a complex life history, developing an operational model of population dynamics of snow crab in an environmentally and spatiotemporally heterogeneous area, the Scotian Shelf of the northwest Atlantic of Canada. We address these difficulties by focussing upon a better and more objective parameterization of snow crab growth based upon surveys of population size structure.*

Keywords - snow crab; Bayesian kernel mixture model; Julia; Turing; growth model; instar

## 1 Introduction

The overall growth patterns snow crab is reasonably well understood. Due to short moult cycles of less than a year in the early stages, growth can be monitored reasonable. However beyond a size range of approximately XX mm, inter-molt periods can be 1 or more years in length. Longevity of snow crab can be from 11 (females) to 15 (males) years. Long term rearing of snow crab is a challenge as living conditions are generally less than ideal or at least different in substantial ways from a natural environment, especially for the larger sized organisms. This can of course create selection biases, such as stunting due to environmental stress (over-crowding, feeding irregularities, water quality, etc.).

There also seems to exist inter-individual and regional-variability in growth patterns due to the interplay environmental (especially bottom temperatures and resource availablity) and potentially genetic factors at large geographic scales, depending upon oceanic currents. Mark-recapture studies can inform such inference for many species, however, in snow crab, due to the loss

of external tags during molts or difficulty and expense of internal tags, this cannot be used effectively.

Observations in more natural settings are also possible by scuba or remote operated vehicle/camera systems. However, such information is quite costly and resource intensive and also susceptible to selection bias in that recaptured or surviving animals tend to be those in better physiological condition than average and so can result in expectations of overly optimistic growth patterns. When routine sampling occurs, a more cost-effective way to establish or corroborate these growth patterns is to decompose observed size frequency distributions. Though there is also inherent size-selection biases in such data due to size-related behavior (habitat preferences) and capture efficiency (net size, speed, depth) that changes with different size and stages, it is still possible to extract some meaningful information of size modes from such data and infer growth patterns.

Given some set of observations of size frequency, subjective "best guesses" (classification by "eye") and implicit reasoning can be used to establish and classify these growth modes. When groups are distinct, this is reasonable. However, observations of size frequencies often demonstrate a mixture of distributions that are heavily overlapping. Even the most state-of-the-art computational algorithms cannot fit such distributions easily. One of the leading more objective approaches to estimate classification using point estimates of latent parameters became available with the development of the Expectation-Maximization (EM) algorithm operating in an Maximum Likelihood framework (Dempster et al 1977; see also the closely related Kalman Filter (Roweis and Ghahramani 1999)) and for distributions in Bayesian frameworks with Variational Inference (Nguyen 2023) and general latent Bayesian Inference. Here we use the latter method, more specifically, **Kernel Mixture Model (KMM)**, using population census-based data to identifying growth stanzas using snow crab data derived from the Maritimes Region of Atlantic Canada.

# 2 Methods

The use of a mixture of distributions has a long history. Pearson (1894) where it was used to identify/classify species of crabs. Holmes (1892) also studied mixture models of wealth disparity. Most numerical methods assigning or classifying data into a cluster or group generally requires the number of such groups to be specified apriori. The excepton being, Infinite Mxture Models (Ghahramani 2011). Fortunately, we have a reasonable understanding of the number of approximate modes of instar carapace widths from visual analysis of size-frequency distributions. This process can be automated by

The finite form of the problem is well known and understood. Implementation is usually with Maximum likelihood using an Expectation-Maximization algorithm (EM; Dempster, Laird, & Rubin, 1977). The solutions to such problems are dependent upon the number of modes chosen, or often the location of the modes, apriori. Many tools exist for estimation:

- https://cran.r-project.org/web/packages/mixtools/vignettes/mixtools.pdf
- https://cran.r-project.org/web/packages/flexmix/vignettes/bootstrapping.pdf
- https://statmath.wu.ac.at/~gruen/BayesMix/bayesmix-intro.pdf

must specify constant node using mcmc Jags

https://cran.r-project.org/web/packages/mclust/vignettes/mclust.html

problems

https://arxiv.org/pdf/2007.04470

https://dr.lib.iastate.edu/server/api/core/bitstreams/333bb46d-c759-4202-8f41-0e921271de53/content

good reviews https://snunnari.github.io/SBE/mclachlan.pdf

https://en.wikipedia.org/wiki/Mixture_model?wprov=sfti1

https://www.sciencedirect.com/topics/medicine-and-dentistry/mixture-model

We can cluster the data using a Bayesian mixture model. The aim of this task is to infer a a latent grouping (hidden structure) from unlabelled data.

Unidimensional Kernel Mixture model with K pre-specified components that cover the space

*alpha* = concentration parameter of 1 (or k, the dimension of the Dirichlet distribution, by the definition used in the topic modelling literature) results in all sets of probabilities being equally likely, i.e., in this case the Dirichlet distribution of dimension k is equivalent to a uniform distribution over a k-1-dimensional simplex. This is not the same as what happens when the concentration parameter tends towards infinity. In the former case, all resulting distributions are equally likely (the distribution over distributions is uniform). In the latter case, only near-uniform distributions are likely (the distribution over distributions is highly peaked around the uniform distribution).

label="observation (i)", ylabel="cluster (k)", legend=false, ) end

https://turing.ml/dev/tutorials/01-gaussian-mixture-model/

# 3 Finite mixtures model

https://turinglang.org/stable/tutorials/01-gaussian-mixture-model/

https://mc-stan.org/users/documentation/case-studies/identifying_mixture_models.html

we want to infer the mixture weights, the parameters \mu_i and the assignment of each datum to a cluster i

standard normal distributions as priors for \mu and a Dirichlet distribution with parameters \alpha_i as prior for w

$$ \begin{aligned} \mu_k &\sim \mathcal{N}(0, 1) \qquad (k = 1,\ldots,K) \ w &\sim \operatorname{Dirichlet}(\alpha_1, \ldots, \alpha_K) \end{aligned}

{} \ {} \

z_i \sim \operatorname{Categorical}(w) \qquad (i = 1,\ldots,N), \

{} \ x_i \sim \mathcal{N}([\mu_{z_i}, \mu_{z_i}]^\mathsf{T}, I) \qquad (i=1,\ldots,N). $$

From simple kernel density representations of size frequency at small area unit scale, determine the magnitudes of modes. This is done as there may be regional and time-dependent changes in modal sizes (year-classes). Sex and maturity status are determined from observation and infered from size-shape changes (carapace width to chela height males or abdominal flap width for females). Inference of modal sizes for each sex-maturity group are determined in order to develop a growth model.

Kernel density estimates of specific components: sex, maturity, year, time of year (season in quarters), region and set (sid).

First we construct kernel density estimates using a bandwidth of 0.025 units on a logarithmic scale. This corresponds to about 4 dx (mm), where dx is the increment width of discretization.

These results are normalized by swept area to provide a density per unit area (1 $\text{km}^{-2}$).

# 4 Results and Discussions

`Coef.  Std. Error      t  Pr(>|t|)  Lower 95%  Upper 95%`

(Intercept) 1.05187 0.0430406 24.44 <1e-04 0.932365 1.17136 instar 0.321978 0.00640426 50.28 <1e-06 0.304196 0.339759

(Intercept) 1.49157 0.216241 6.90 0.0917 -1.25602 4.23917 instar 0.268652 0.0221858 12.11 0.0525 -0.0132459 0.550549

(Intercept) 1.10376 0.0259636 42.51 <1e-08 1.04237 1.16515 instar 0.308079 0.00308857 99.75 <1e-11 0.300775 0.315382

(Intercept) 0.335281 0.114697 2.92 0.2098 -1.12208 1.79264 instar 0.363371 0.00953601 38.11 0.0167 0.242204 0.484537

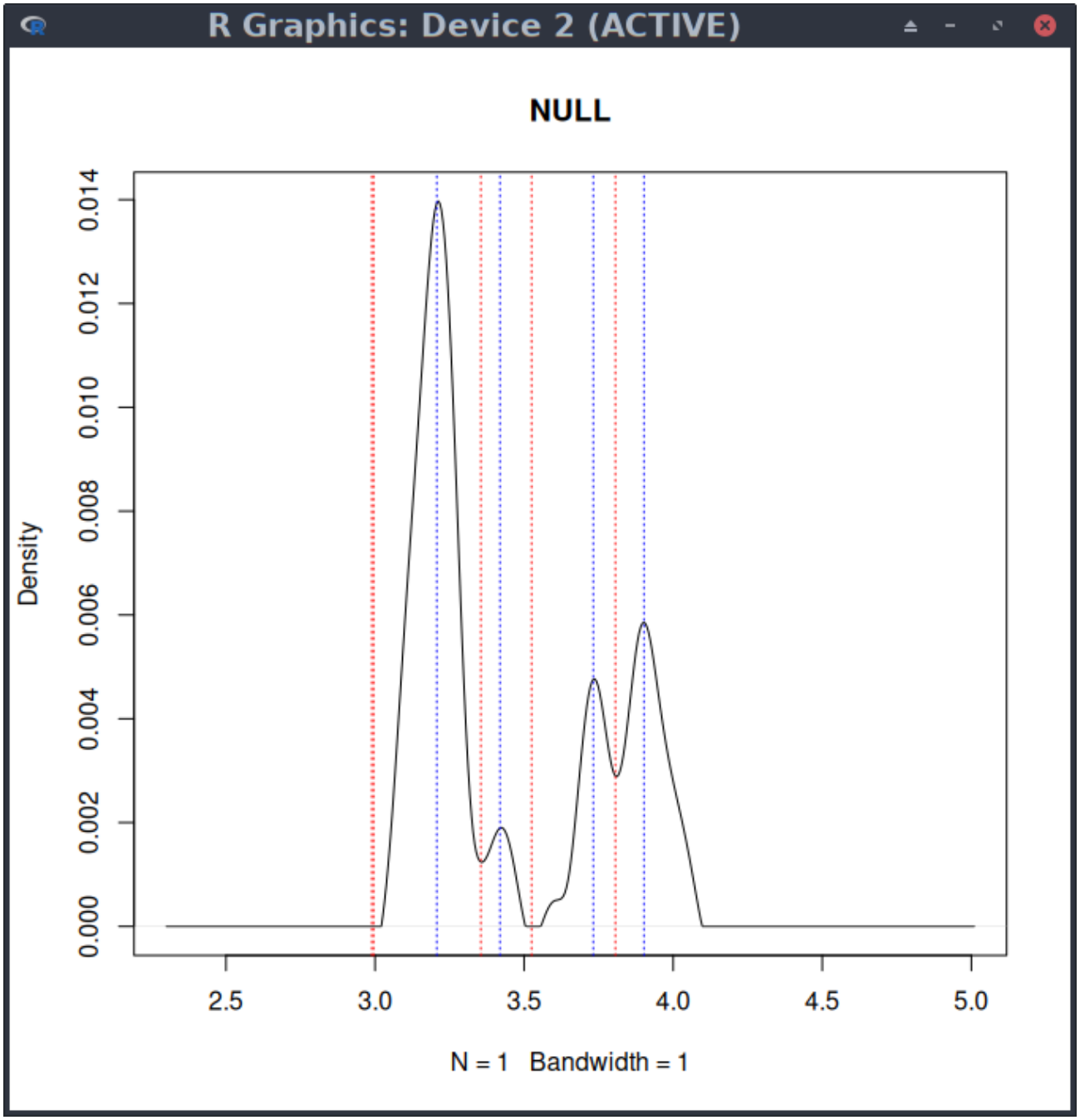
R Graphics: Device 2 (ACTIVE)
NULL
Density
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
2.5
3.0
3.5
4.0
4.5
5.0
N = 1 Bandwidth = 1

**Fig. 1:** *image*

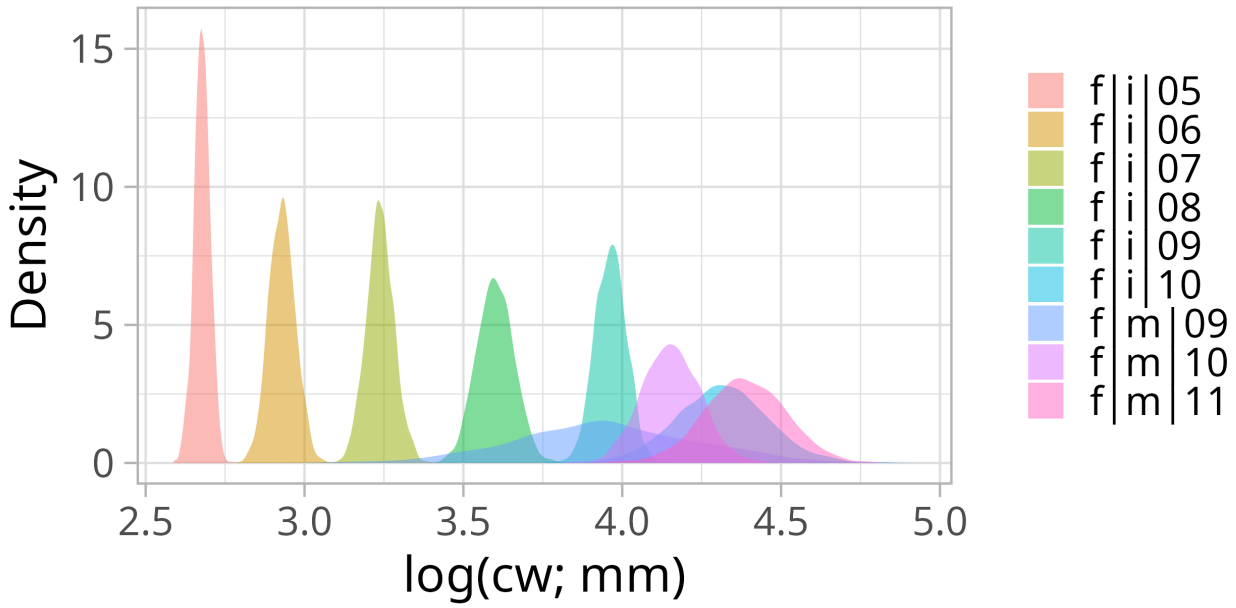


**Fig. 2:** *image*

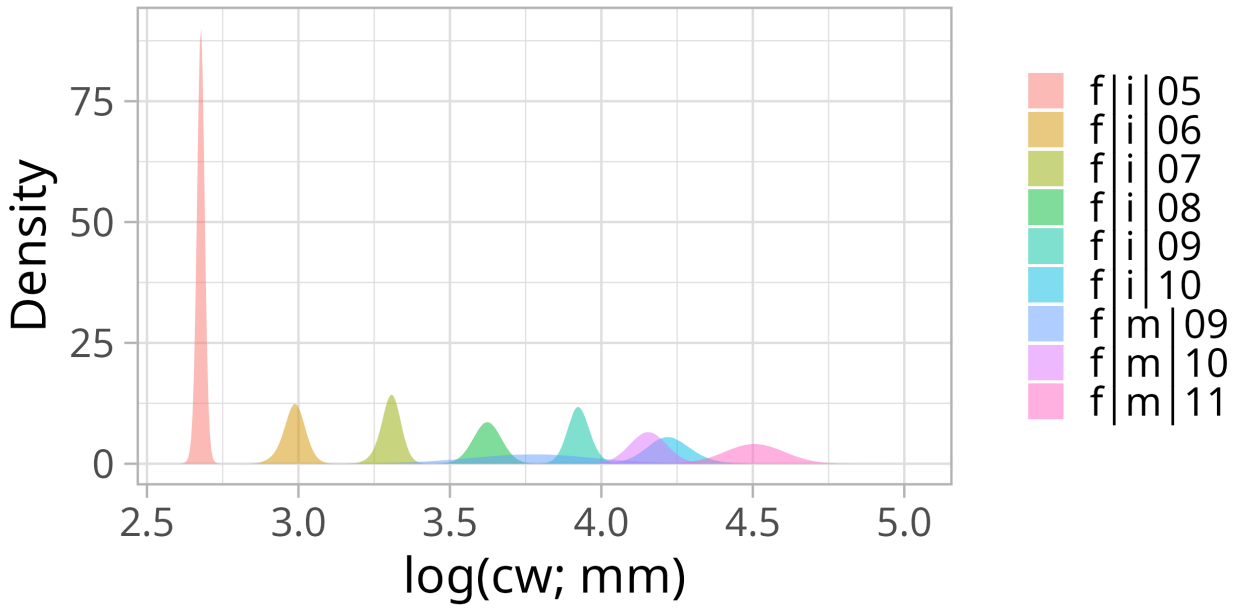


**Fig. 3:** *image*

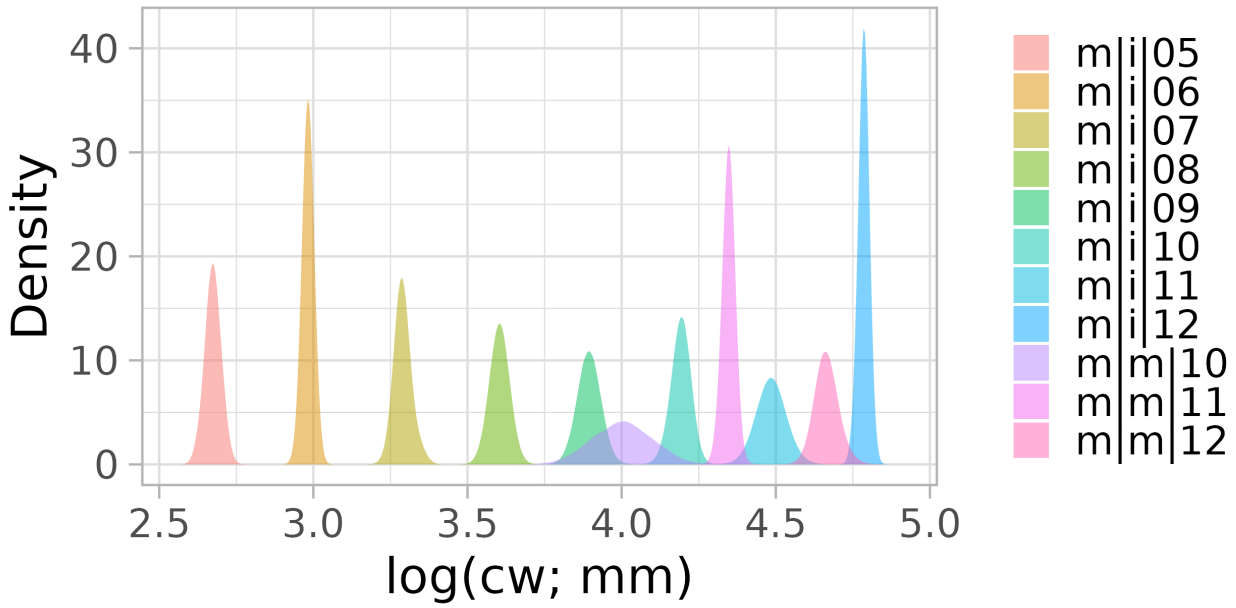


**Fig. 4:** *image*

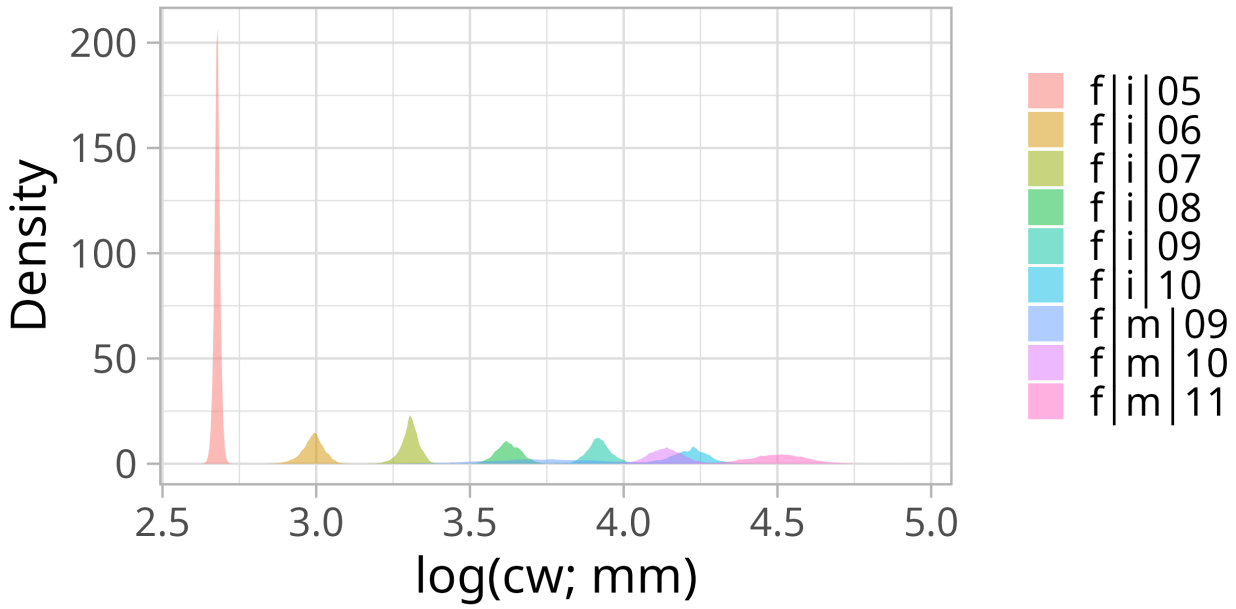


**Fig. 5:** *image*

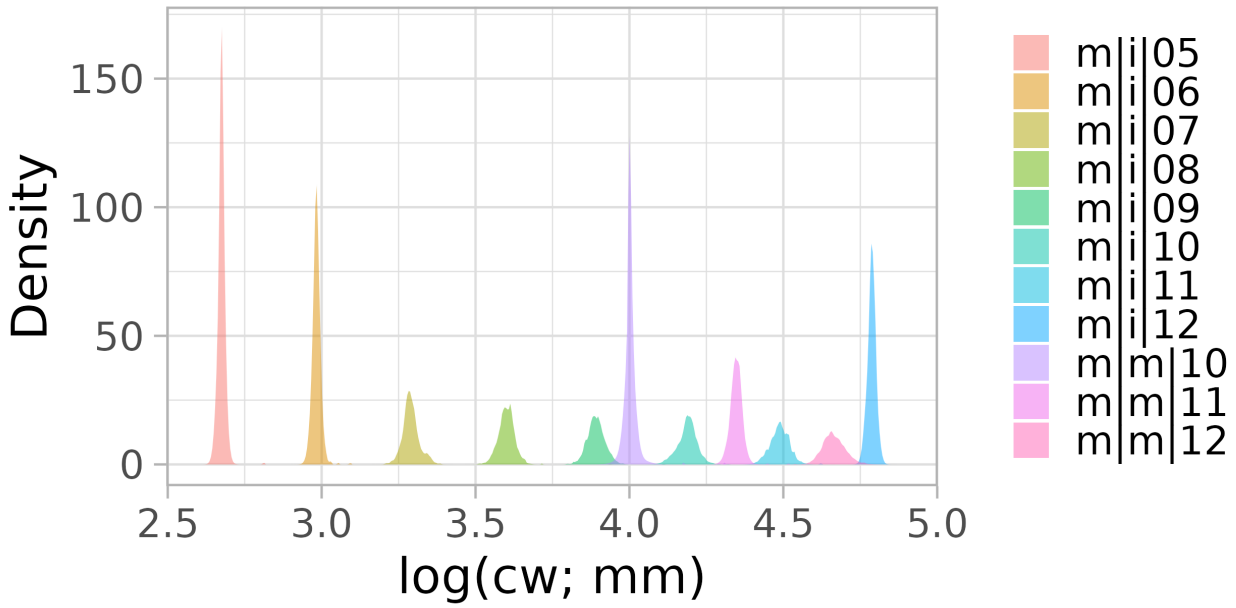


**Fig. 6:** *image*

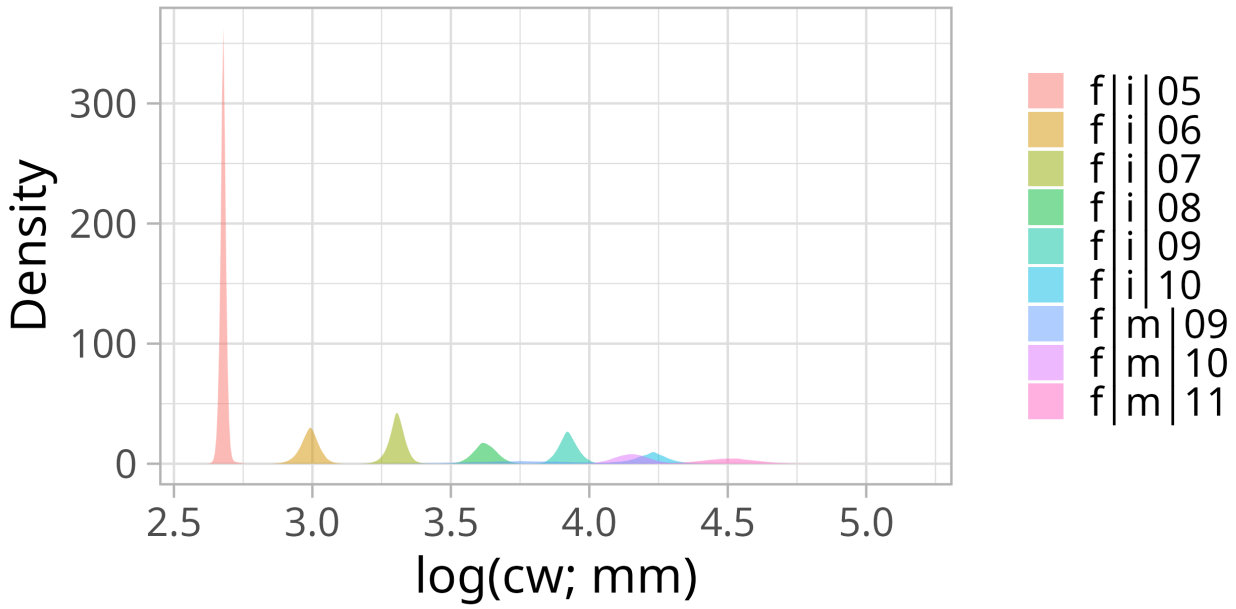


**Fig. 7:** *image*

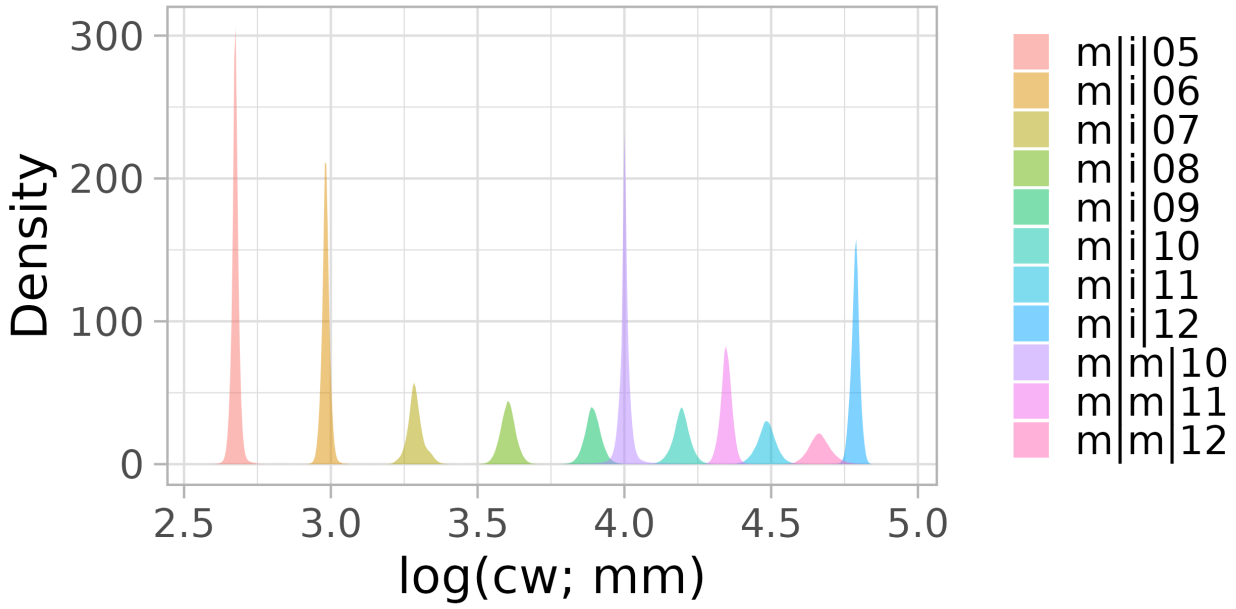
300
200
100
0
Density
2.5
3.0
3.5
4.0
4.5
5.0
log(cw; mm)
m|i|05
m|i|06
m|i|07
m|i|08
m|i|09
m|i|10
m|i|11
m|i|12
m|m|10
m|m|11
m|m|12

**Fig. 8:** *image*

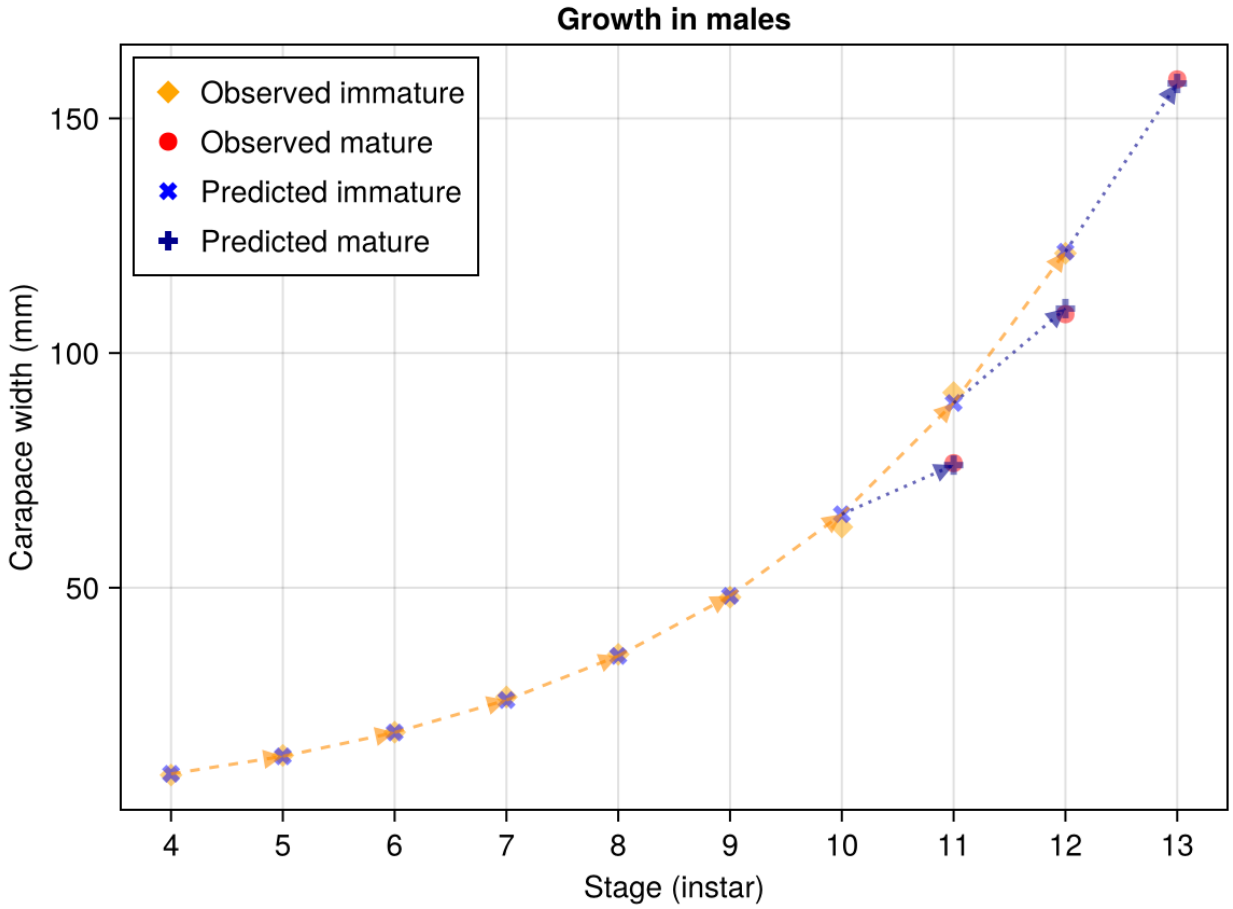
Growth in males
Observed immature
Observed mature
Predicted immature
Predicted mature
Carapace width (mm)
150
100
50
4
5
6
7
8
9
10
11
12
13
Stage (instar)

**Fig. 9:** *image*

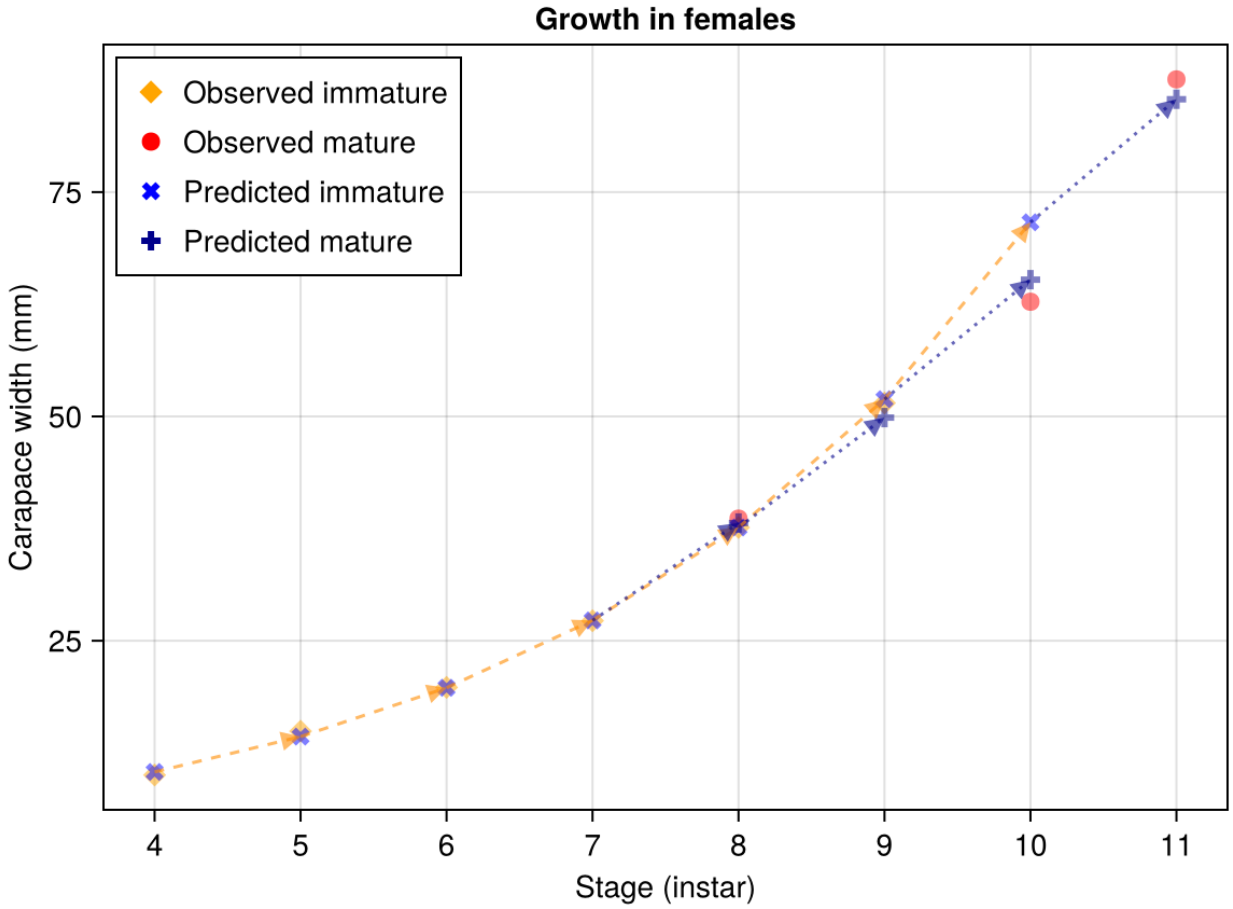
Growth in females
Observed immature
Observed mature
Predicted immature
Predicted mature
Carapace width (mm)
75
50
25
4
5
6
7
8
9
10
11
Stage (instar)

**Fig. 10:** *image*