Comparing Parametric and Nonparametric Models

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Motivation

Two approaches to estimate pmfs from data

1. Empirical pmf (nonparametric model)

2. Parametric model based on a predefined distribution

Which one is better?

Goal

Learn how to evaluate models in realistic situations

Free throws

Goal: Model streaks of consecutive free throws

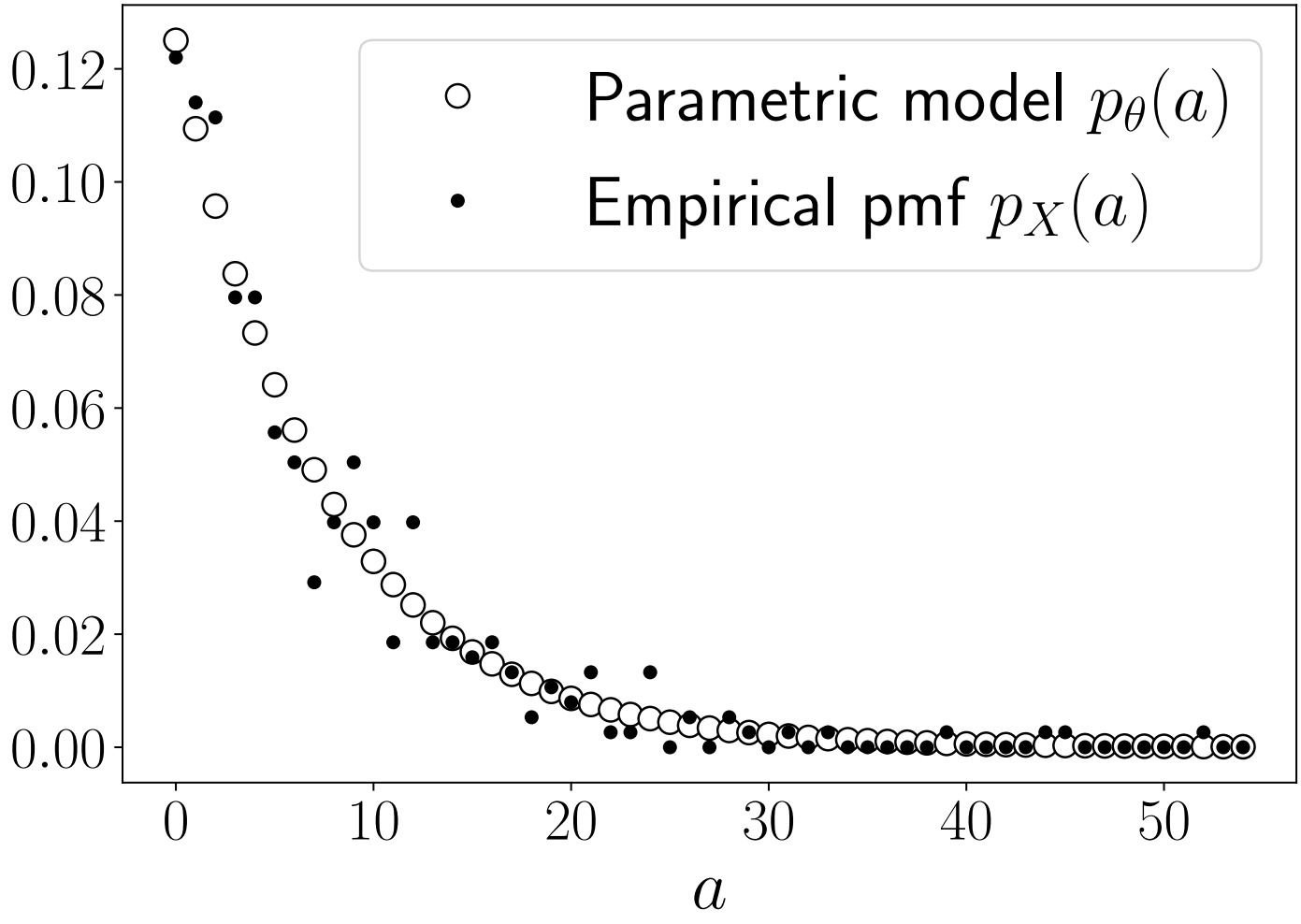
Data: 377 streaks from 3,015 free throws shot by Kevin Durant in the NBA

Nonparametric model: Empirical pmf

Parametric model:

$$p_\theta(s) = \theta^s(1-\theta)$$

Free throws


0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
50
a
Parametric model $p_\theta(a)$
Empirical pmf $p_X(a)$

Evaluation strategy 1

Which one provides the best fit to the observed data?

Empirical pmf provides perfect fit

Sounds too good to be true... What is the problem?

You can't use the same data to fit the model and evaluate it!

Evaluation strategy 2

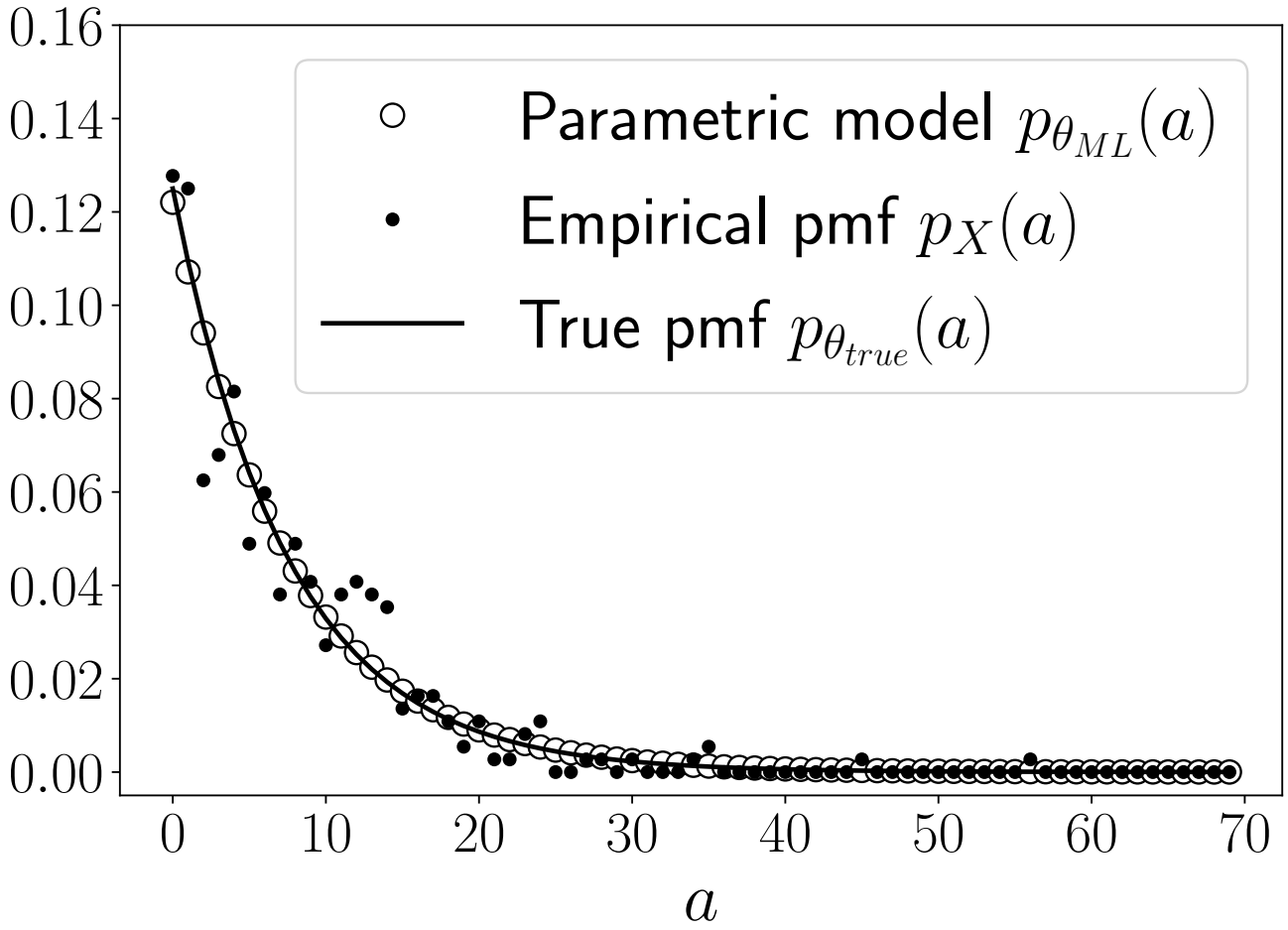
Which one provides the best fit on simulated data assuming parametric model holds?

We simulate 3,015 i.i.d. free throws from

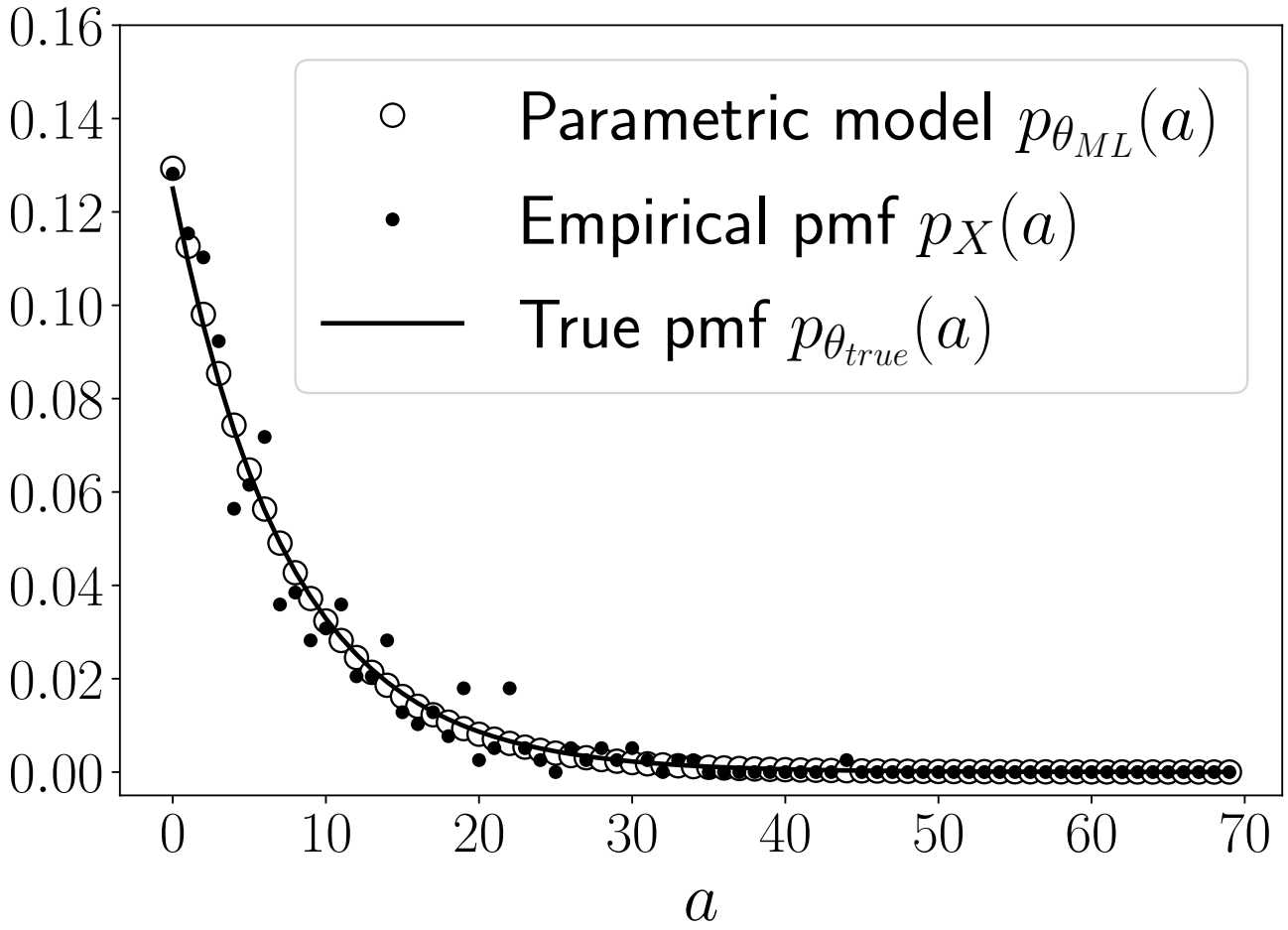
$$p_\theta(s) = \theta_{\text{true}}^s(1 - \theta_{\text{true}})$$

with $\theta_{\text{true}} := 0.875$

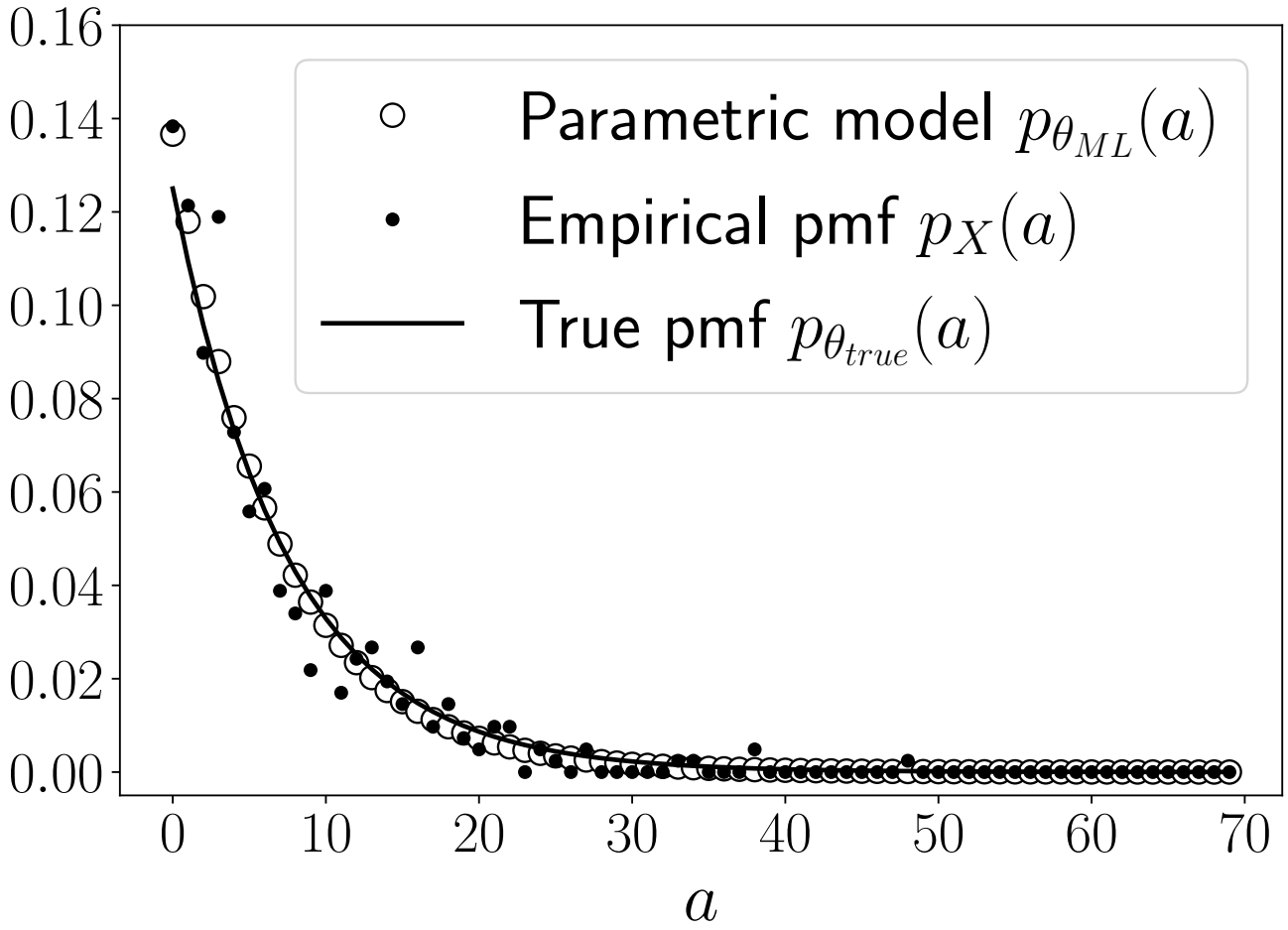
$\theta_{\text{ML}} = 0.873$



$\theta_{\text{ML}} = 0.879$



$\theta_{\text{ML}} = 0.867$



Case closed?

Should we conclude that parametric model is better? No!

Real data are not generated by parametric model!

Then how do we evaluate? Use real held-out data

Training and test set

Training set: Data used to fit (*train*) the model

Test set: Data used to evaluate (*test*) the model

First 3,015 free throws used for training, following 3,015 for test

Evaluation metric

Root mean square error (RMSE) between estimated pmf p_{est} and empirical test pmf p_{test}

$$\text{RMSE}(p_{\text{est}}) := \sqrt{\frac{1}{L}\sum_{\ell=0}^{L}(p_{\text{est}}(\ell) - p_{\text{test}}(\ell))^2}$$

Nonparametric model

Not really *nonparametric* because it requires estimating entries of pmf

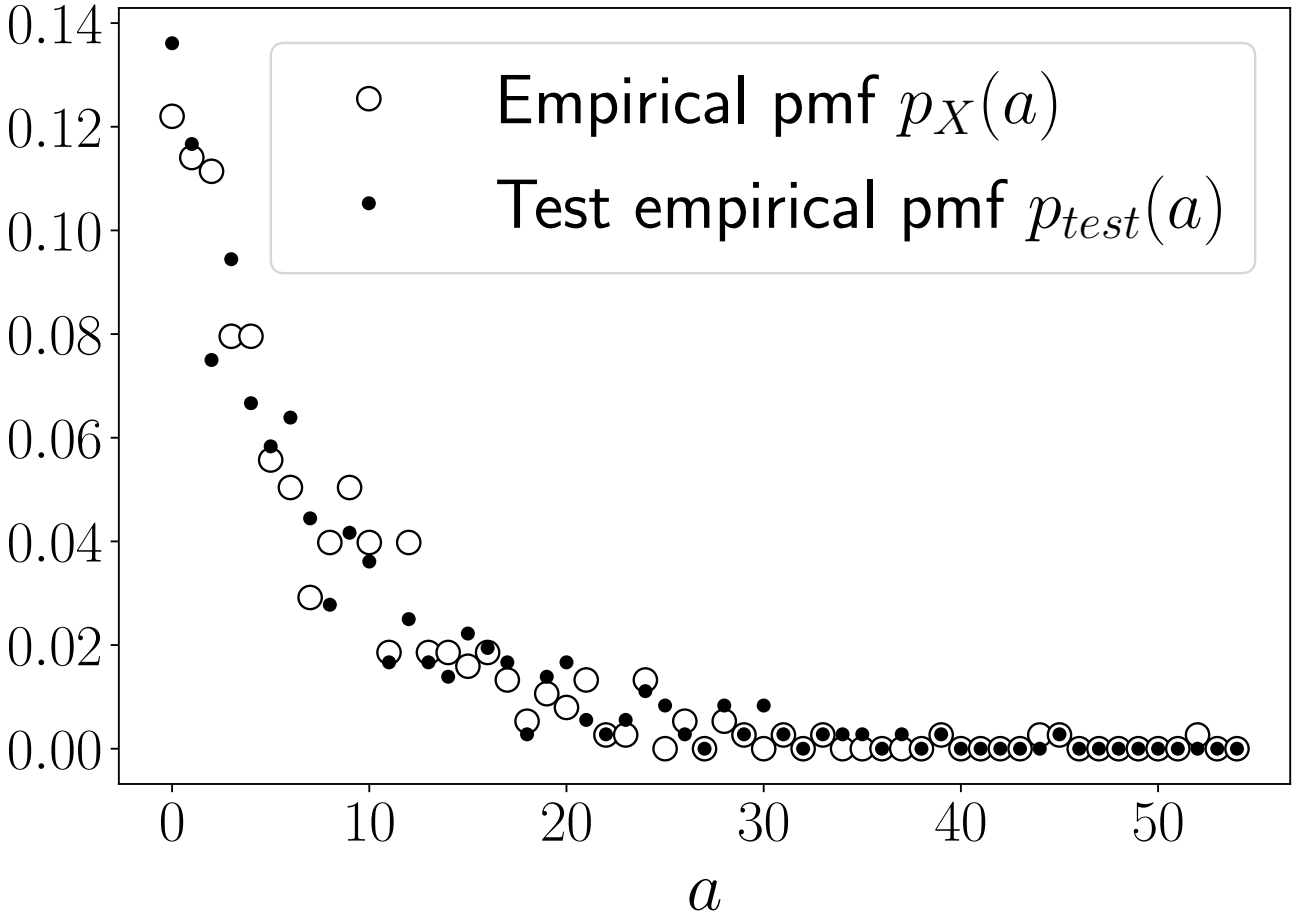
Number of estimated parameters: 55 (compare to 1 for parametric model!)

Good news: A lot of flexibility to learn from data

Bad news: Can overfit noise in training data

Nonparametric model

Test RMSE $= 7.67 \cdot 10^{-3}$



Parametric model

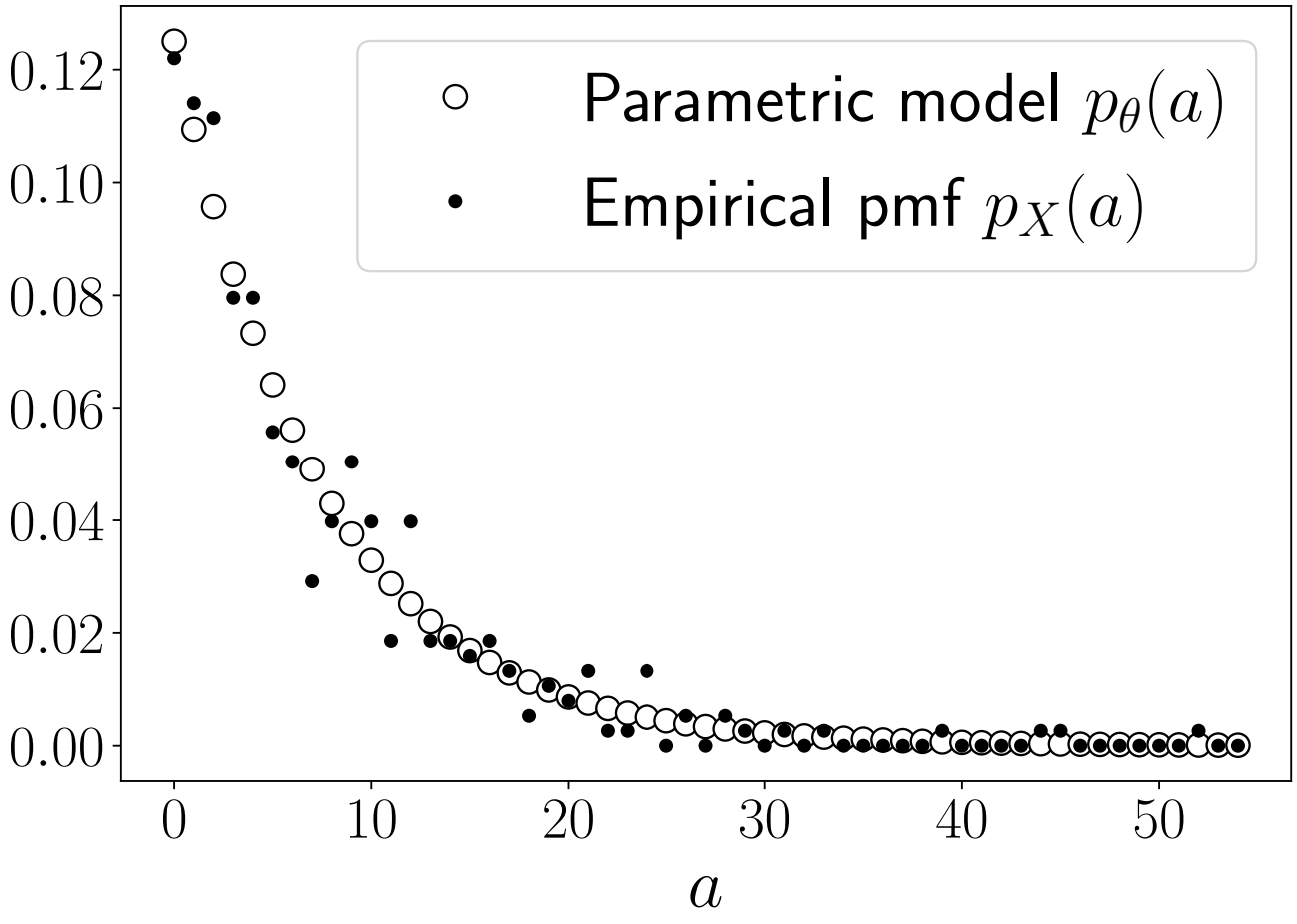
Predefined shape constrains model

Good news: Cannot overfit easily

Bad news: Less flexibility to fit signal, can underfit signal in training data

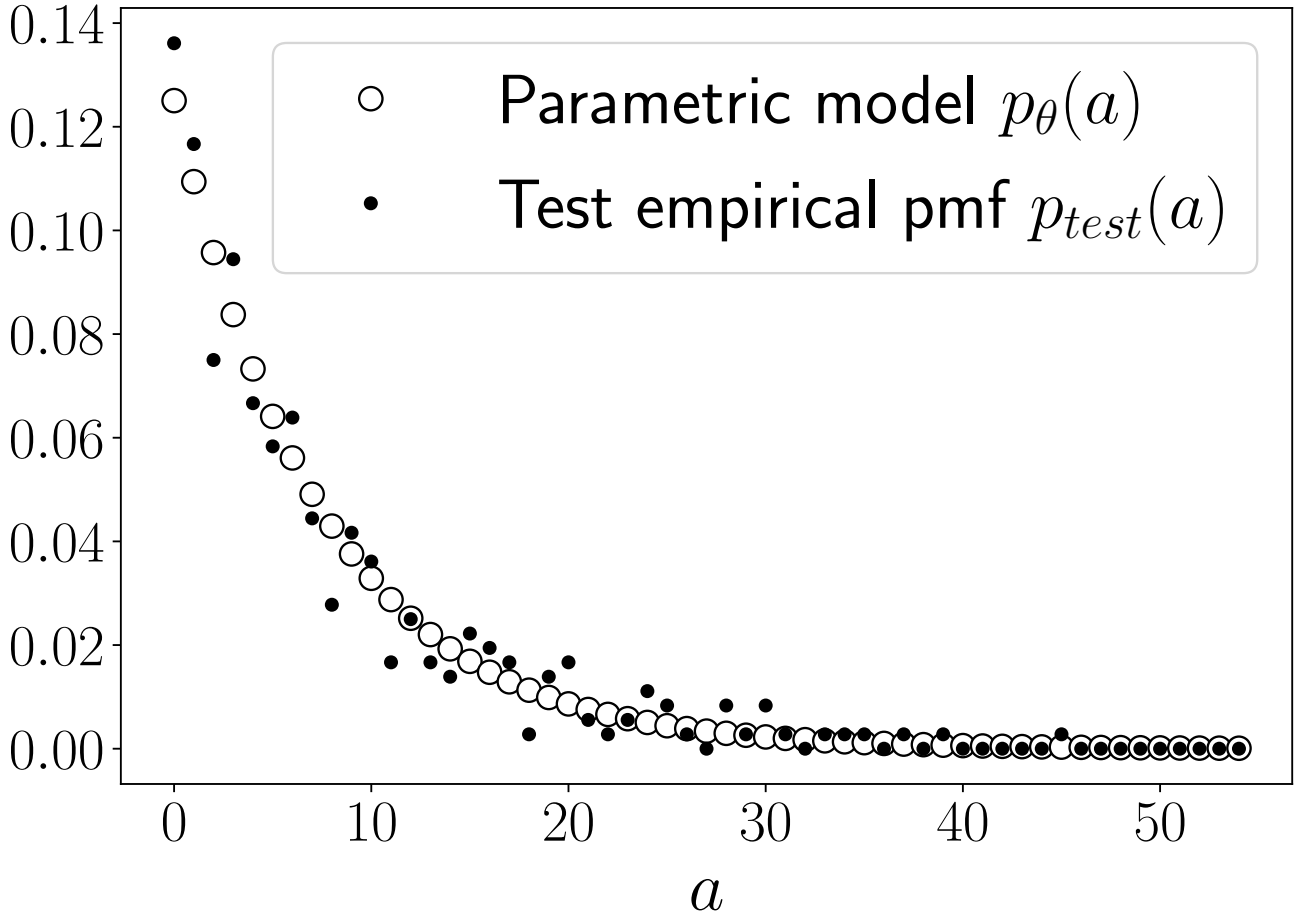
Parametric model

Training RMSE $= 5.46 \cdot 10^{-3}$



Parametric model

Test RMSE $= 5.61 \cdot 10^{-3}$



Parametric or nonparametric?

Not a lot of data, but good knowledge of data-generating process?

Use parametric model

Little knowledge about data-generating process, but a lot of data?

Use nonparametric model

Call center in bank

Goal: Model number of calls between 6 am and 7 am on weekdays

Training set: Calls from January-June 1999

Test set: Calls from July-December 1999

Parametric model

Assumptions:

1. Calls are independent
2. Probability of a call in period of small length t is λt
3. Probability of more calls in when $t \to 0$ is negligible

Parametric model? Poisson!

Likelihood

$$\begin{aligned}\mathcal{L}_X(\lambda) &= \prod_{i=1}^{n} p_{\tilde{x}}(x_i) \\ &= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\end{aligned}$$

$$\log \mathcal{L}_X(\lambda) = \sum_{i=1}^{n} \left(x_i \log \lambda - \lambda - \log\left(x_i!\right)\right)$$

Maximum likelihood

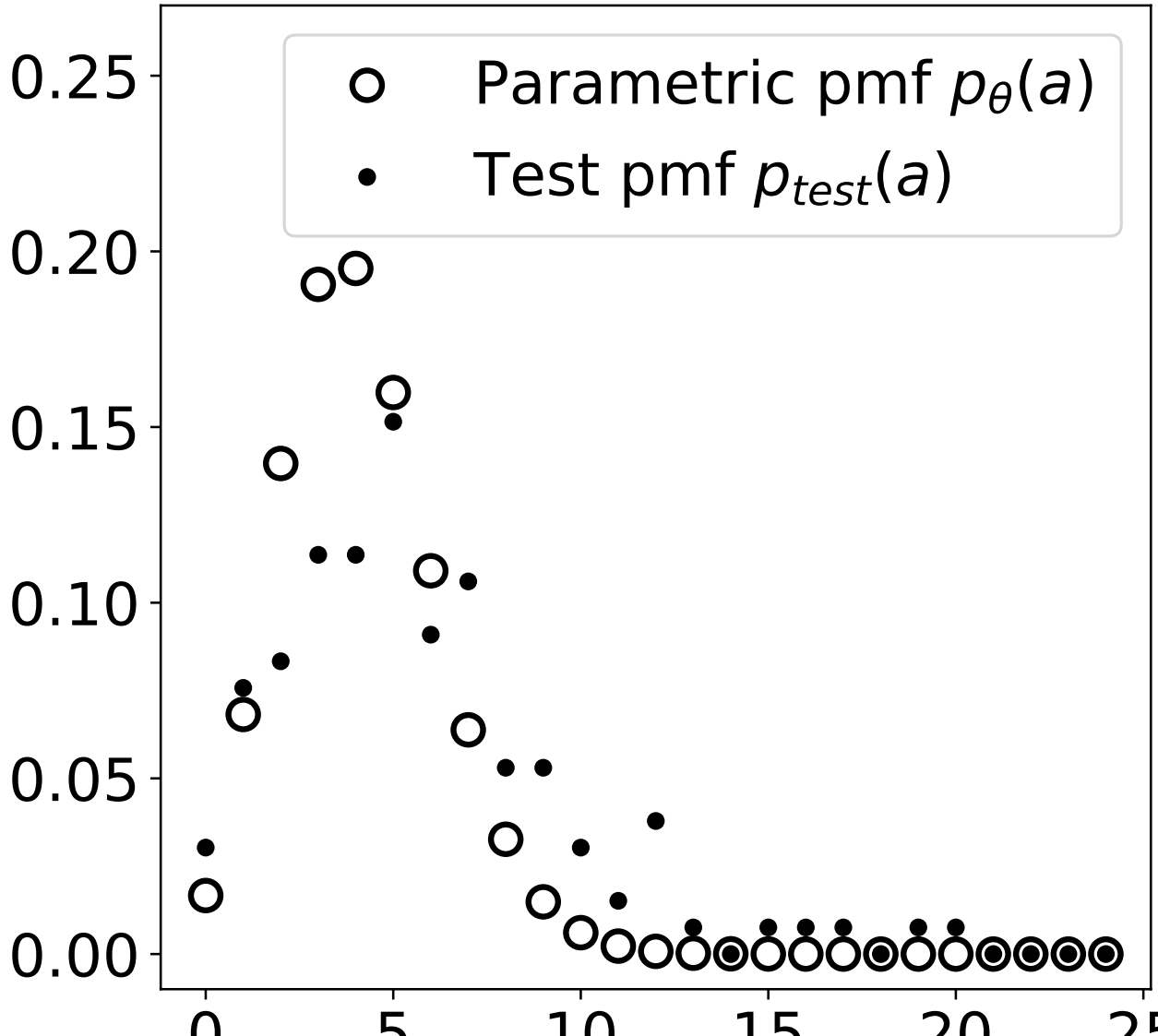
$$\log \mathcal{L}_X(\lambda) = \sum_{i=1}^{n} (x_i \log \lambda - \lambda - \log(x_i!))$$

$$\frac{\mathrm{d} \log \mathcal{L}_X(\lambda)}{\mathrm{d}\lambda} = \sum_{i=1}^{n} \frac{x_i}{\lambda} - 1 = 0 \quad \rightarrow \quad \lambda_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\frac{\mathrm{d}^2 \log \mathcal{L}_X(\lambda)}{\mathrm{d}\lambda^2} = -\sum_{i=1}^{n} \frac{x_i}{\lambda^2} < 0$$

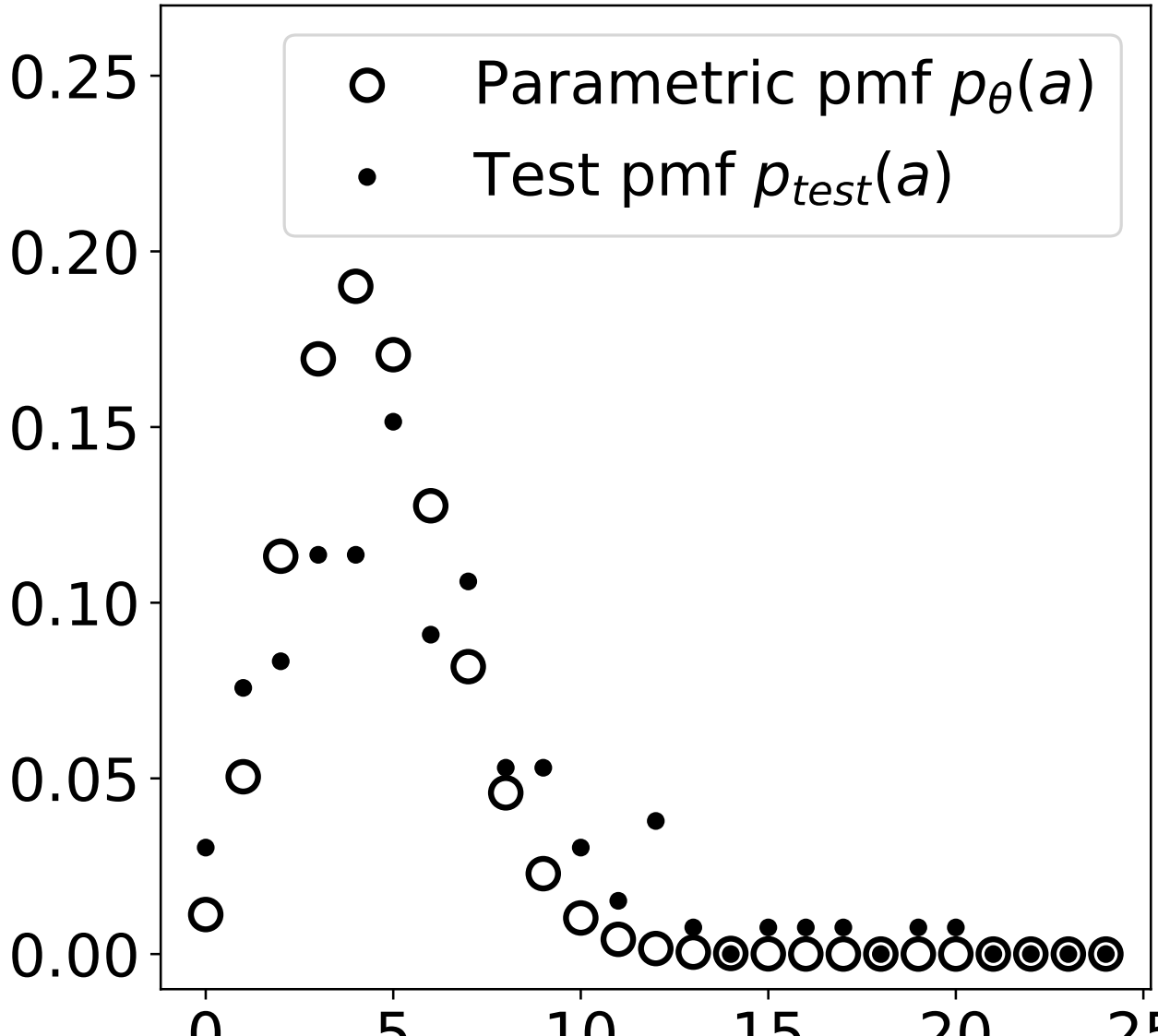
Parametric model

Training months: January



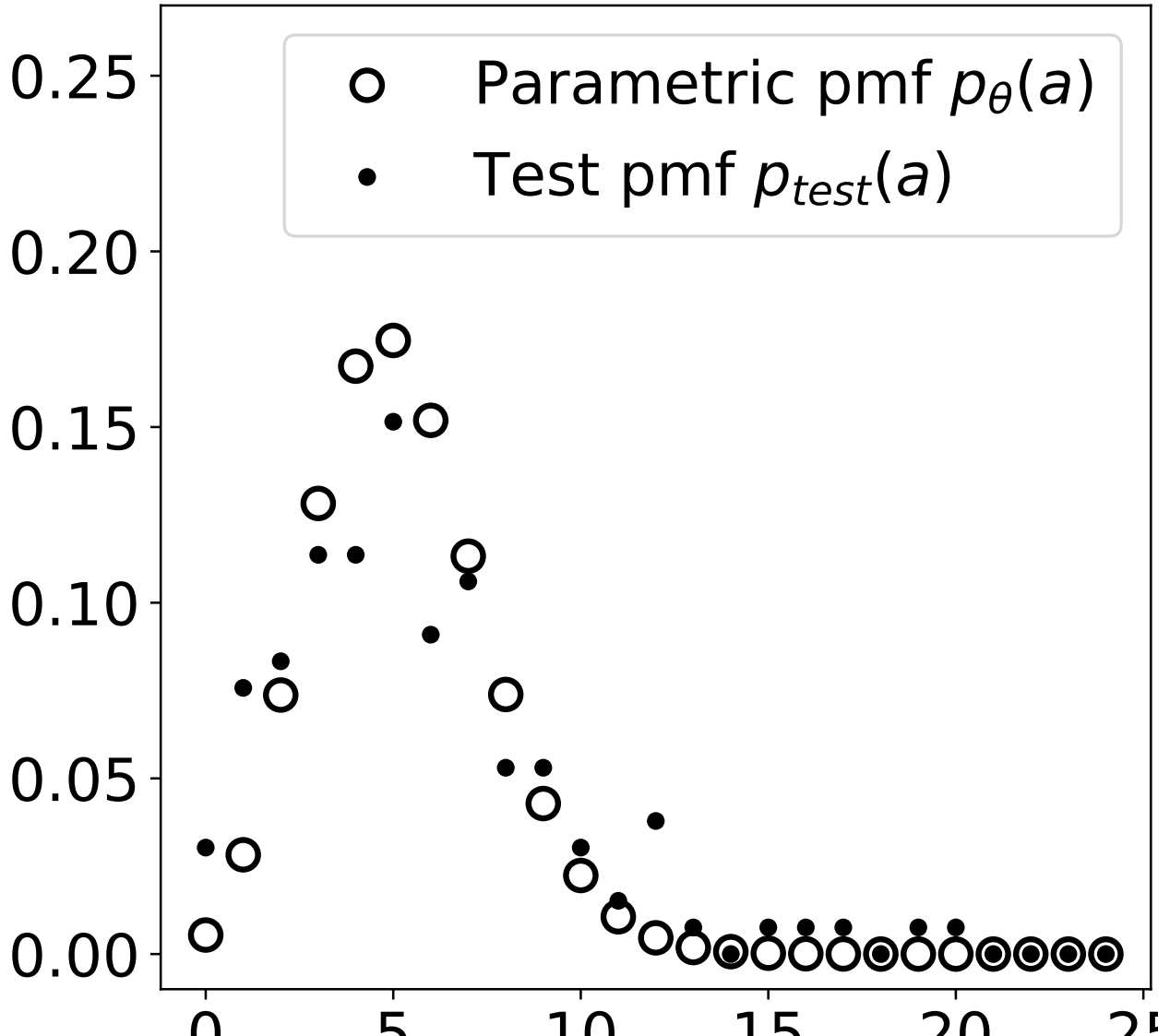
Parametric model

Training months: January-March



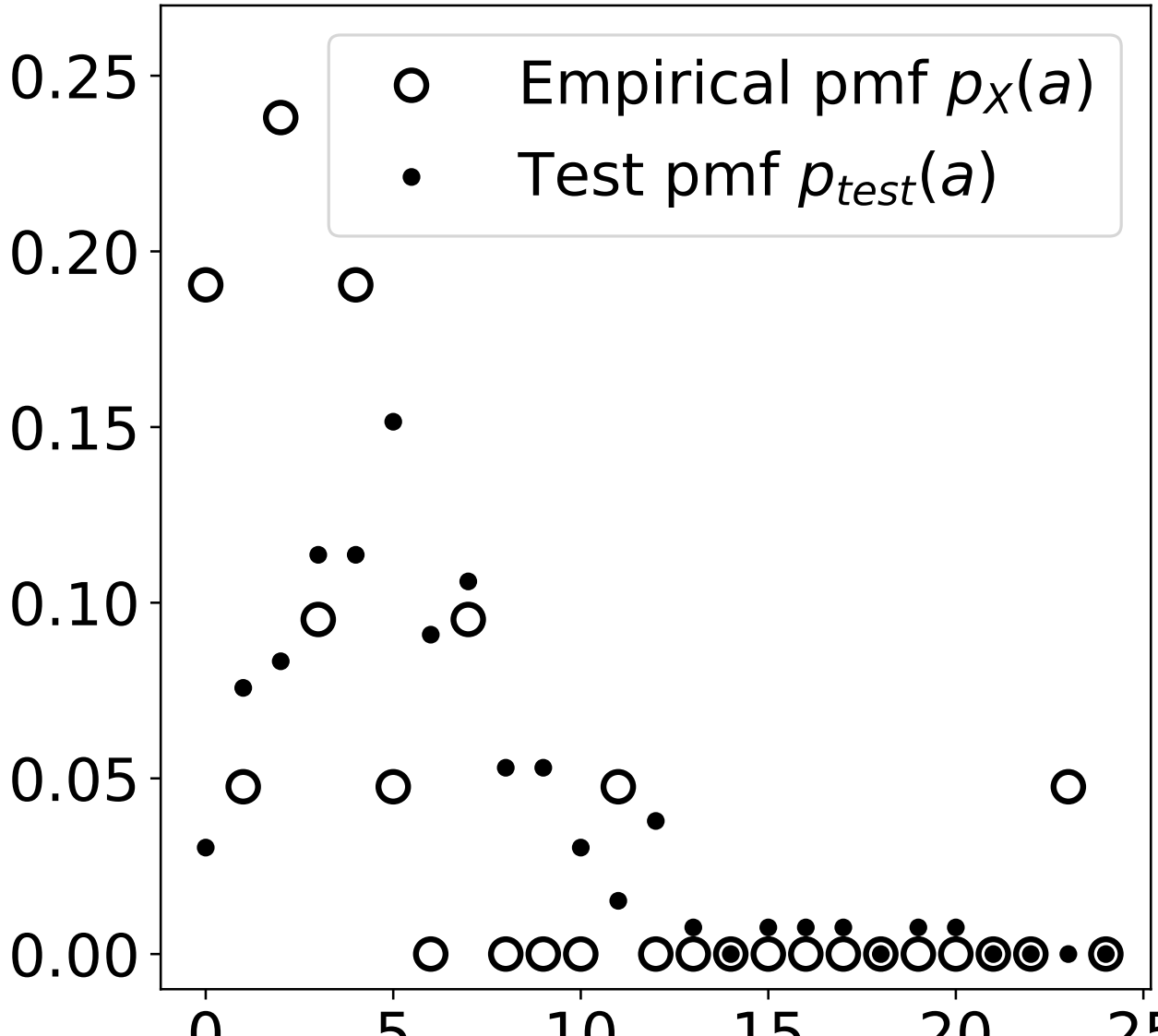
Parametric model

Training months: January-June



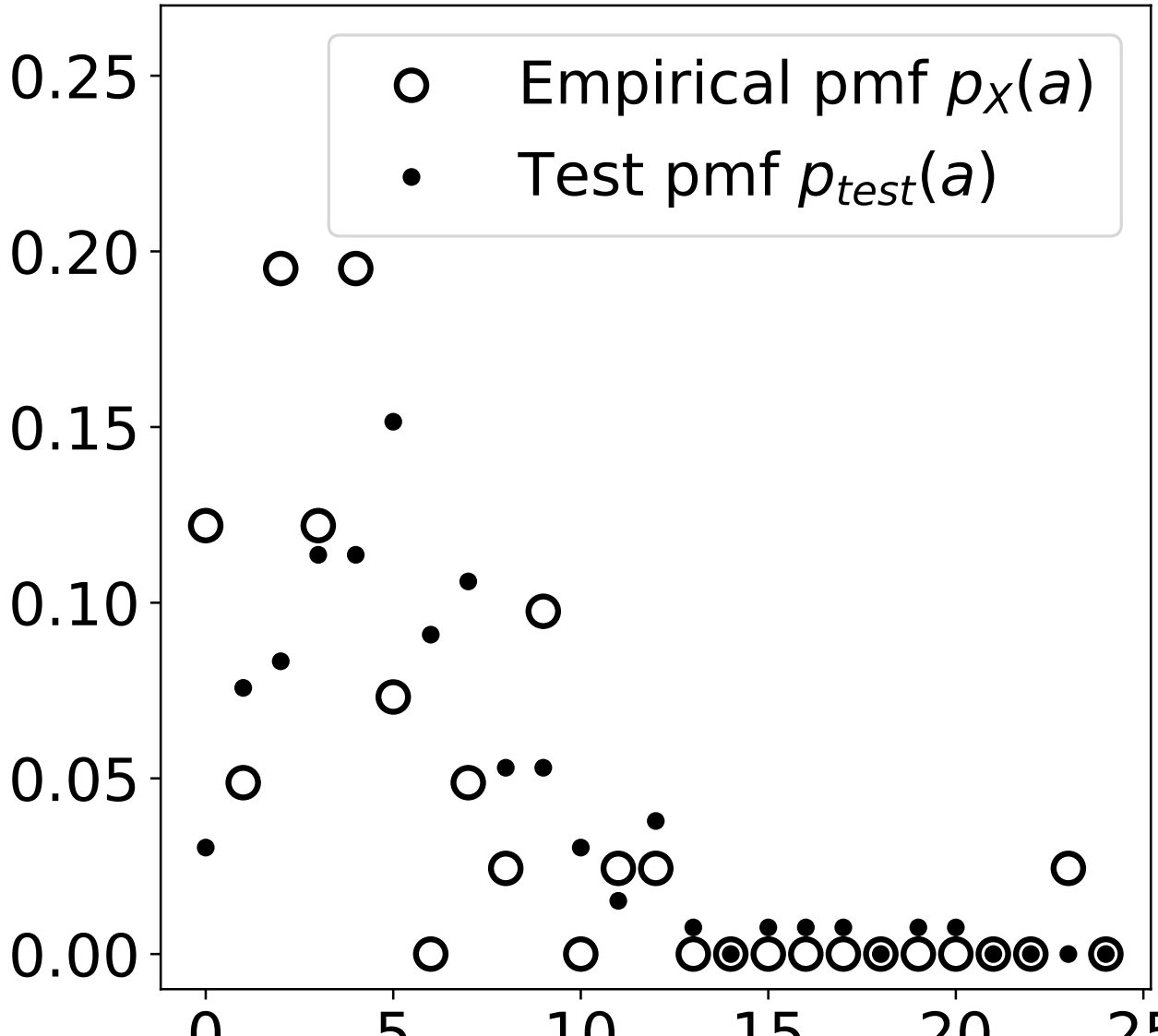
Nonparametric model

Training months: January



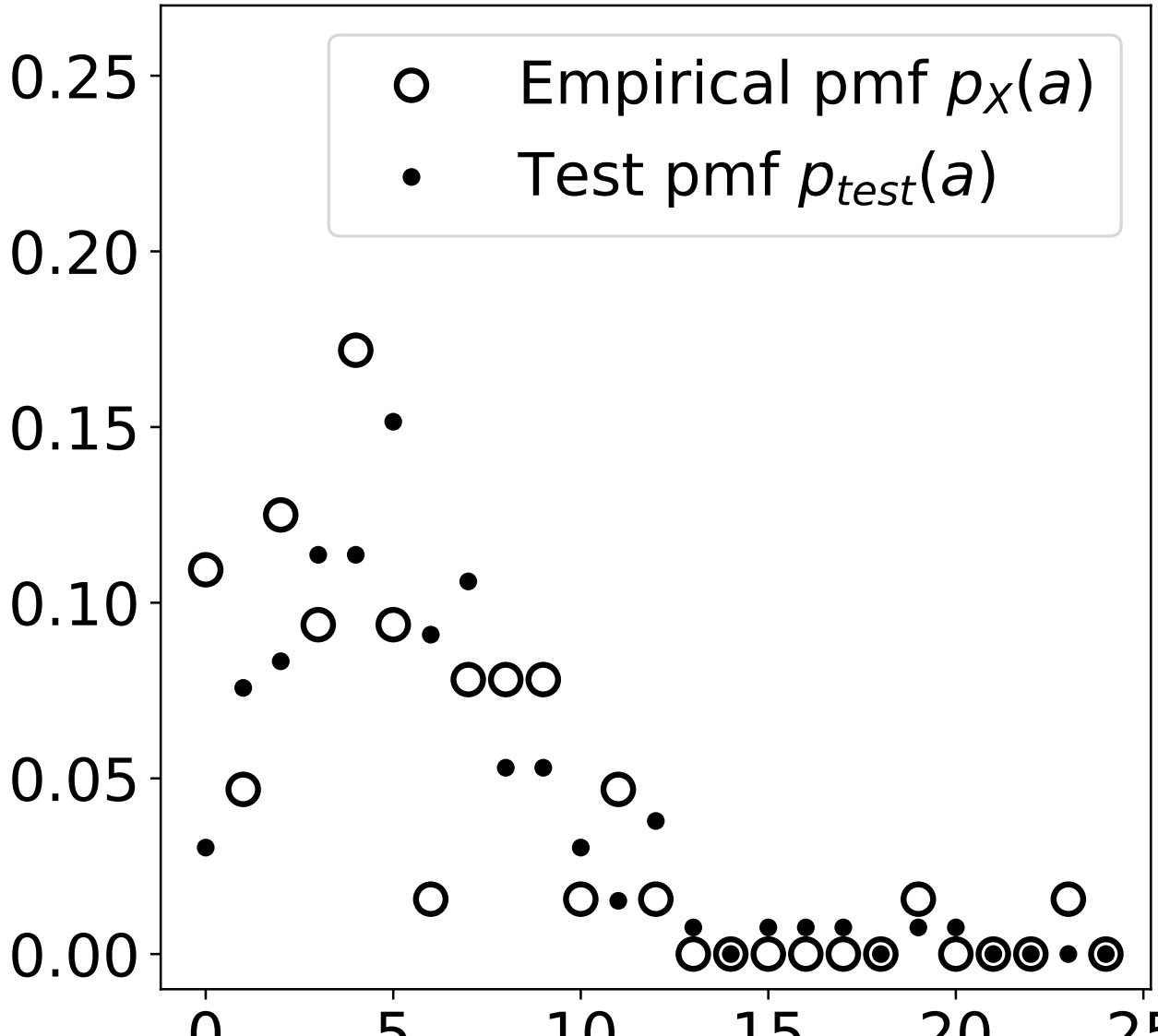
Nonparametric model

Training months: January-March

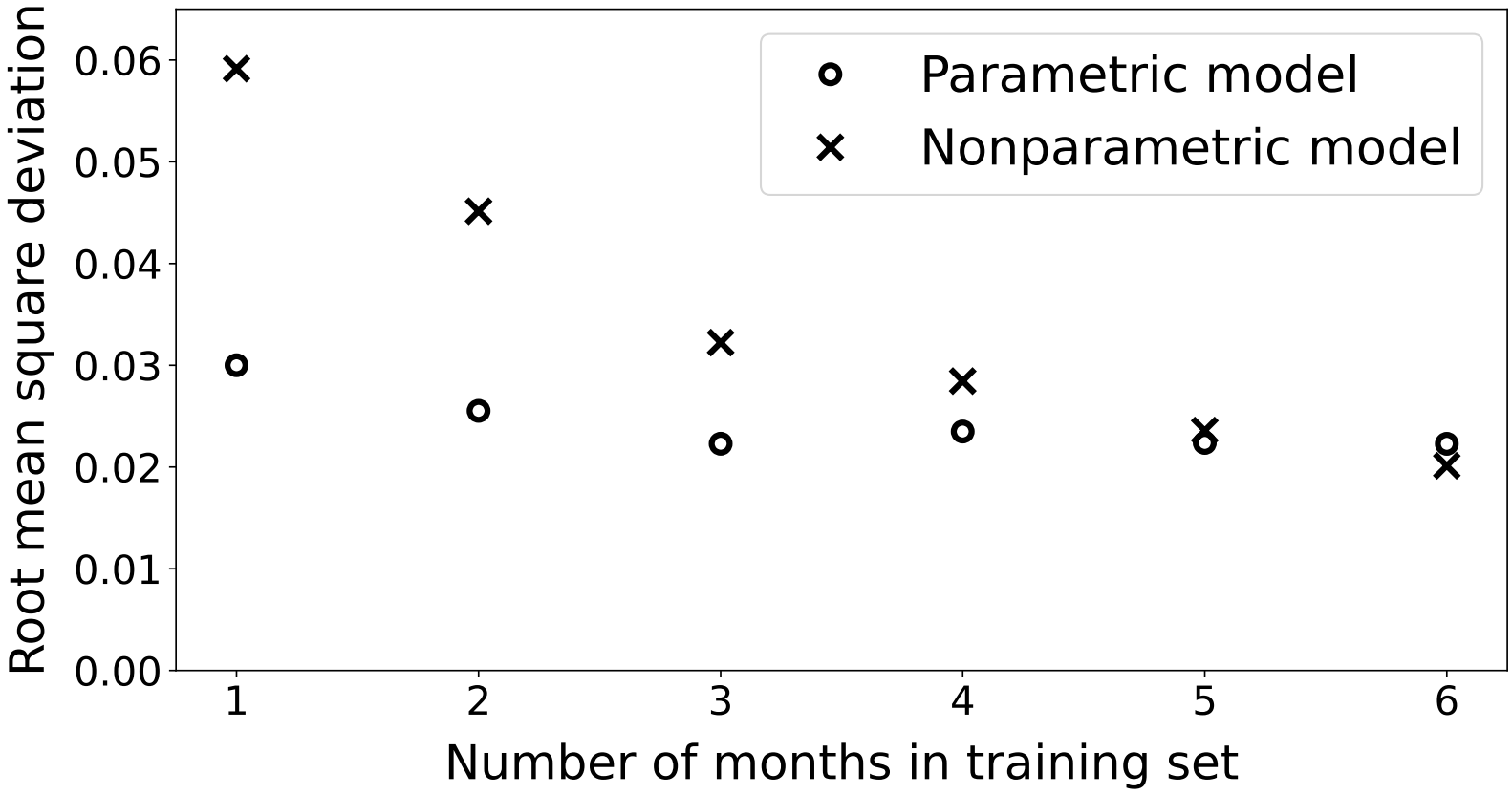


Nonparametric model

Training months: January-June



Test RMSE


Root mean square deviation
0.00
0.01
0.02
0.03
0.04
0.05
0.06
1
2
3
4
5
6
Number of months in training set
Parametric model
Nonparametric model

What have we learned?

How to evaluate models in practice

Advantages / disadvantages of parametric and nonparametric models