Thoughts on Model Selection

Chad M. Schafer
Department of Statistics & Data Science
Carnegie Mellon University

April 2023

Contributions

Thanks to

Lucas Kania,

Mikael Kuusela,

Nick Wardle, and

Larry Wasserman,

for their contributions to this talk.

Overview

“Statistician’s view on model selection”

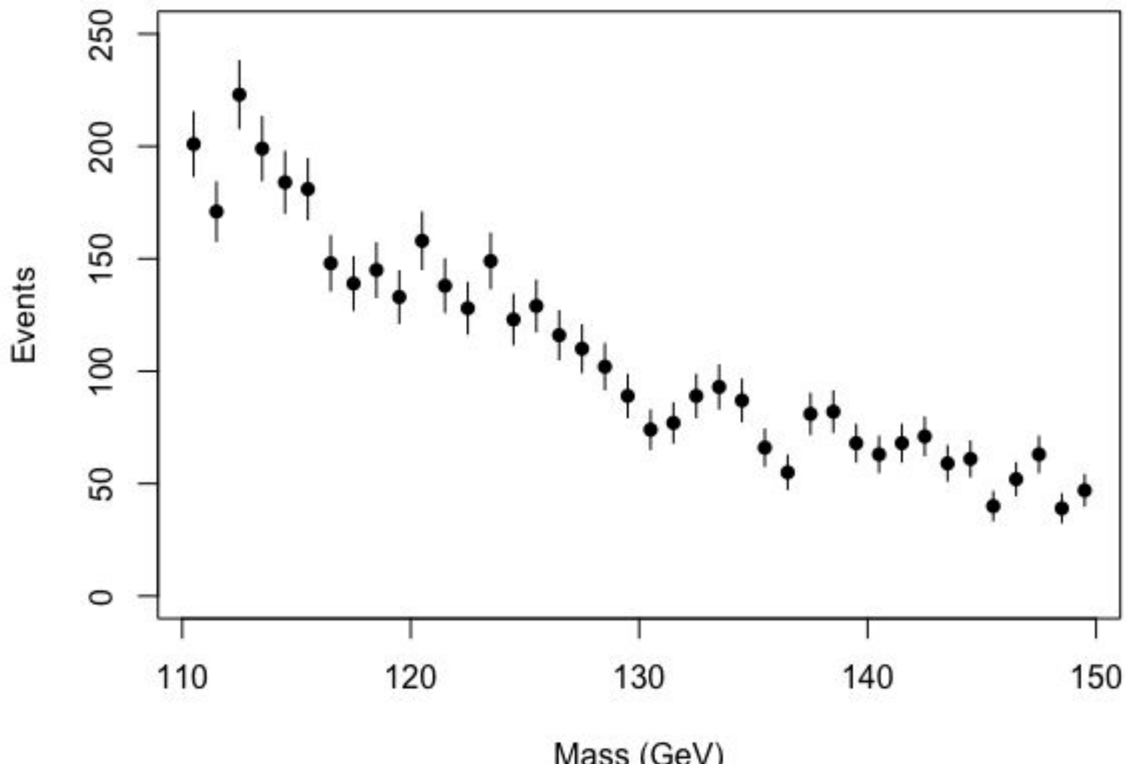
Very large topic

Touch on a few, relevant (I hope), concepts and ideas

- parametric versus nonparametric models
- AIC
- discrete profiling
- semiparametric models
- method of sieves
- model averaging

Focus on the specific problem of **background model selection** and the implications for **post-model selection inference**

Overview



Simulated data, courtesy of Nick Wardle

Models

General form:

$$\lambda_i = \eta(m_i) + \theta N(125, 1.19)$$

$$X_i \sim \text{Poisson}(\lambda_i)$$

Consider polynomial background models, estimate via maximum likelihood

Models

General form:

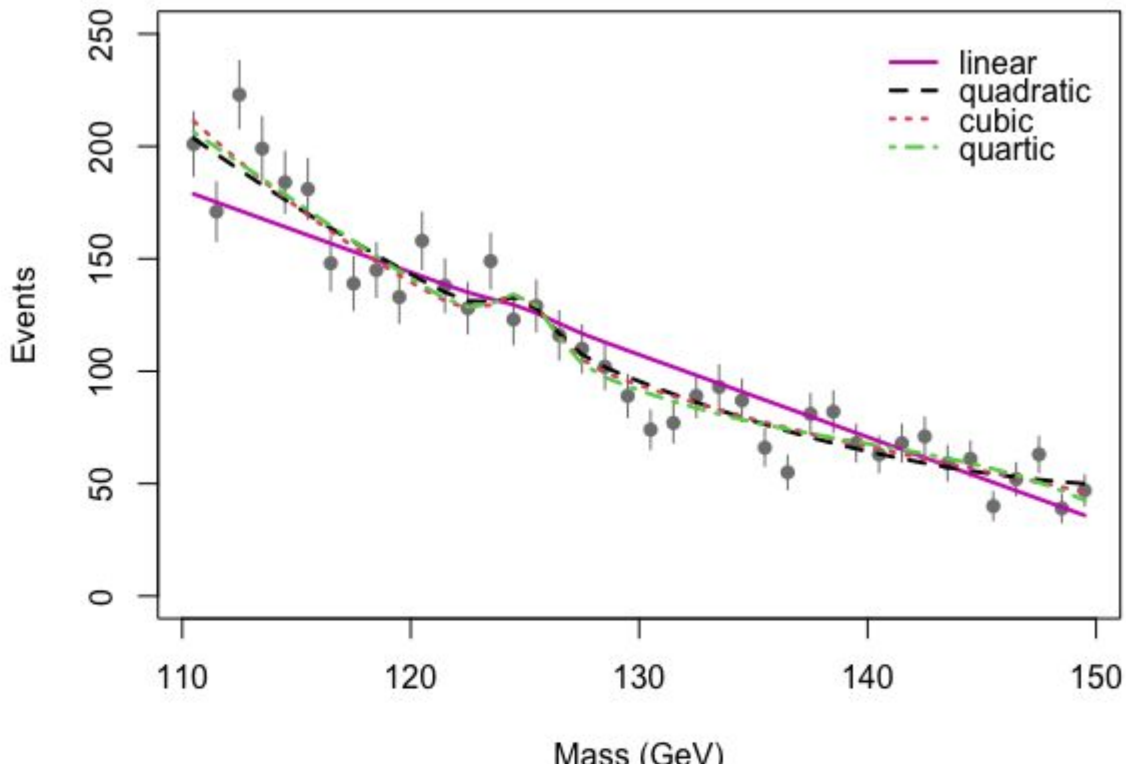
Background **Signal**

$$\lambda_i = \eta(m_i) + \theta N(125, 1.19)$$

$$X_i \sim \mathrm{Poisson}(\lambda_i)$$

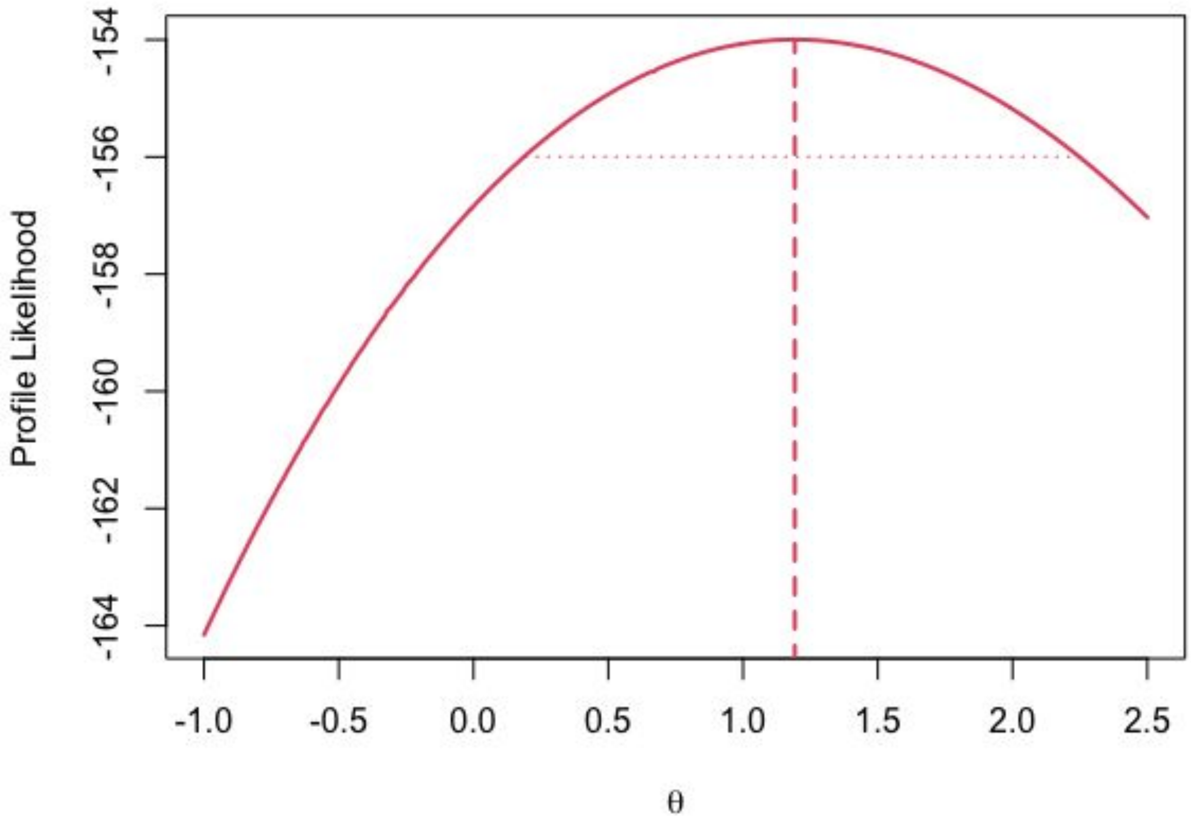
Consider polynomial background models, estimate via maximum likelihood

Models



Simulated data, courtesy of Nick Wardle

Models



Profile likelihood for the quartic case

Models

Model	MLE of θ	95% CI for θ	Log likelihood at MLE	Number of Parameters	AIC
Linear	0.13	(-0.75, 1.12)	-174.79	3	355.58
Quadratic	0.84	(-0.14, 1.80)	-155.85	4	319.70
Cubic	1.14	(0.2, 2.10)	-154.81	5	319.62
Quartic	1.21	(0.23, 2.17)	-154.02	6	320.05

$$\text{AIC} = -2 \times \text{log likelihood at max} + 2 \times \text{number of parameters}$$

Models

Model	MLE of θ	95% CI for θ	Log likelihood at MLE	Number of Parameters	AIC
Linear	0.13	(-0.75, 1.12)	-174.79	3	355.58
Quadratic	0.84	(-0.14, 1.80)	-155.85	4	319.70
Cubic	1.14	(0.2, 2.10)	-154.81	5	319.62
Quartic	1.21	(0.23, 2.17)	-154.02	6	320.05
Quintic	1.18	(0.29, 2.18)	-153.84	7	321.67
Sextic	1.24	(0.33, 2.19)	-153.83	8	323.65
Septic	1.25	(0.39, 2,30)	-153.34	9	324.69

Parametric versus Nonparametric

Is this a **parametric** or a **nonparametric** approach?

The process described above for choosing the background model is nonparametric.

The order of the model is serving as a **smoothing parameter**.

AIC is a widely used approach to choosing its value.

More data = less “smoothing”

Semiparametric Models

Defines a de facto s**emiparametric approach**, where complexity of model is limited by ensuring consistency of estimation of θ

Semiparametric inference:

Parameter of interest: θ, lies in a Euclidean space

Nuisance parameter: η, lies in a more general space, denote it **Ɲ**

Semiparametric Models

Murphy and van der Vaart (2000), "On Profile Likelihood":

"We show that semiparametric profile likelihoods, where the ***nuisance parameter has been profiled out****, behave like ordinary likelihoods in that they have a quadratic expansion. In this expansion the score function and the Fisher information are replaced by the* ***efficient score function*** *and* ***efficient Fisher information****.*

"

Semiparametric Models

Murphy and van der Vaart (2000), “On Profile Likelihood”:

“We show that semiparametric profile likelihoods, where the ***nuisance parameter has been profiled out****, behave like ordinary likelihoods in that they have a quadratic expansion. In this expansion the score function and the Fisher information are replaced by the* ***efficient score function*** *and* ***efficient Fisher information****. The expansion may be used, among others, to prove the asymptotic normality of the maximum likelihood estimator, to derive the asymptotic chi-squared distribution of the log-likelihood ratio statistic, and to prove the consistency of the observed information as an estimator of the inverse of the asymptotic variance.”*

Semiparametric Models

The **score function for θ**

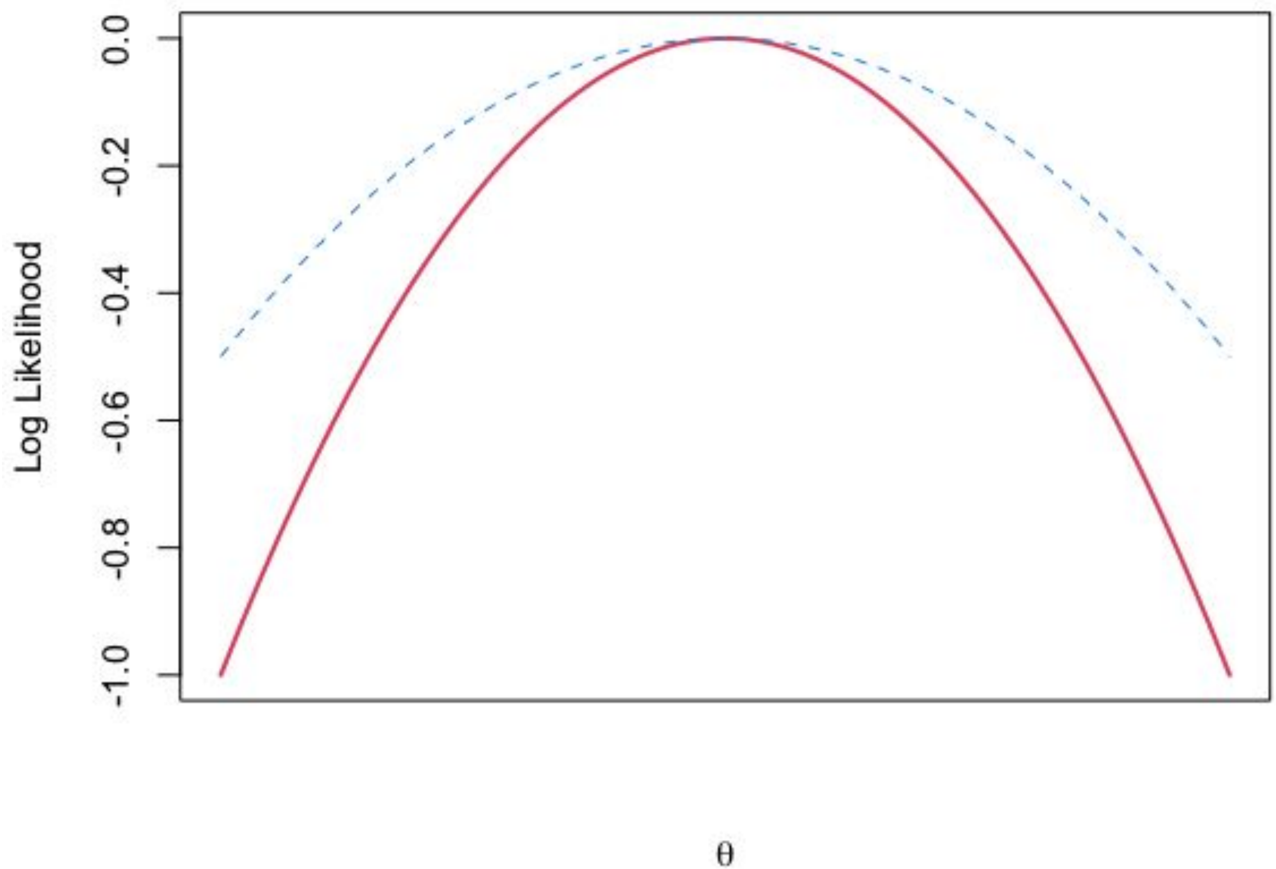
$$l_{\theta_0,\eta_0} = \frac{\partial}{\partial\theta} \log L(\theta,\eta)\Big|_{\theta_0,\eta_0}$$

The **efficient score function for θ**

$$\tilde{l}_{\theta_0,\eta_0} = l_{\theta_0,\eta_0} - \Pi_{\theta_0,\eta_0} l_{\theta_0,\eta_0}$$

where Π projects onto the space of score functions for η, finding the **least favorable model in Ɲ**

Semiparametric Models



Semiparametric Models

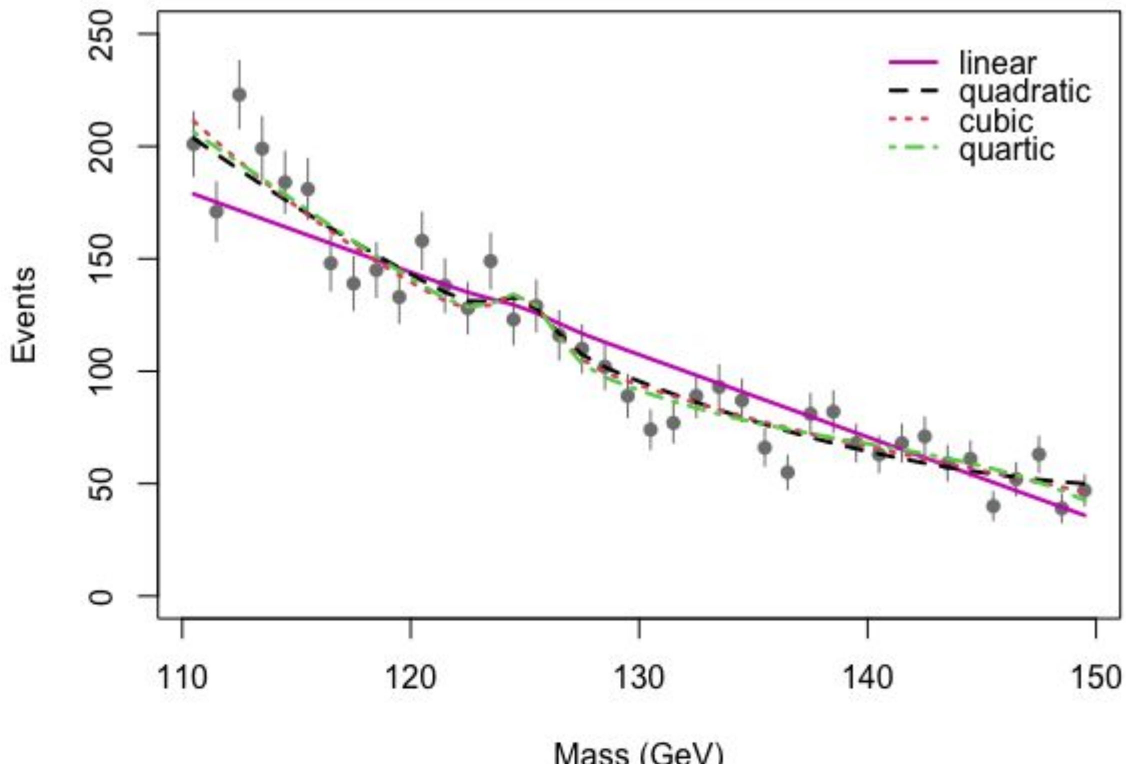
The **efficient Fisher Information**

$$\mathrm{Var}(\tilde{l}_{\theta_0,\eta_0}) = \tilde{\mathcal{I}}_{\theta_0,\eta_0}$$

The MLE for θ is approximately normal with covariance matrix

$$\tilde{\mathcal{I}}^{-1}_{\widehat{\theta},\widehat{\eta}}$$

Semiparametric Models



Simulated data, courtesy of Nick Wardle

Semiparametric Models

Important Question:

What controls the complexity of the background?

I.e., what limits N?

Could utilize physical constraints

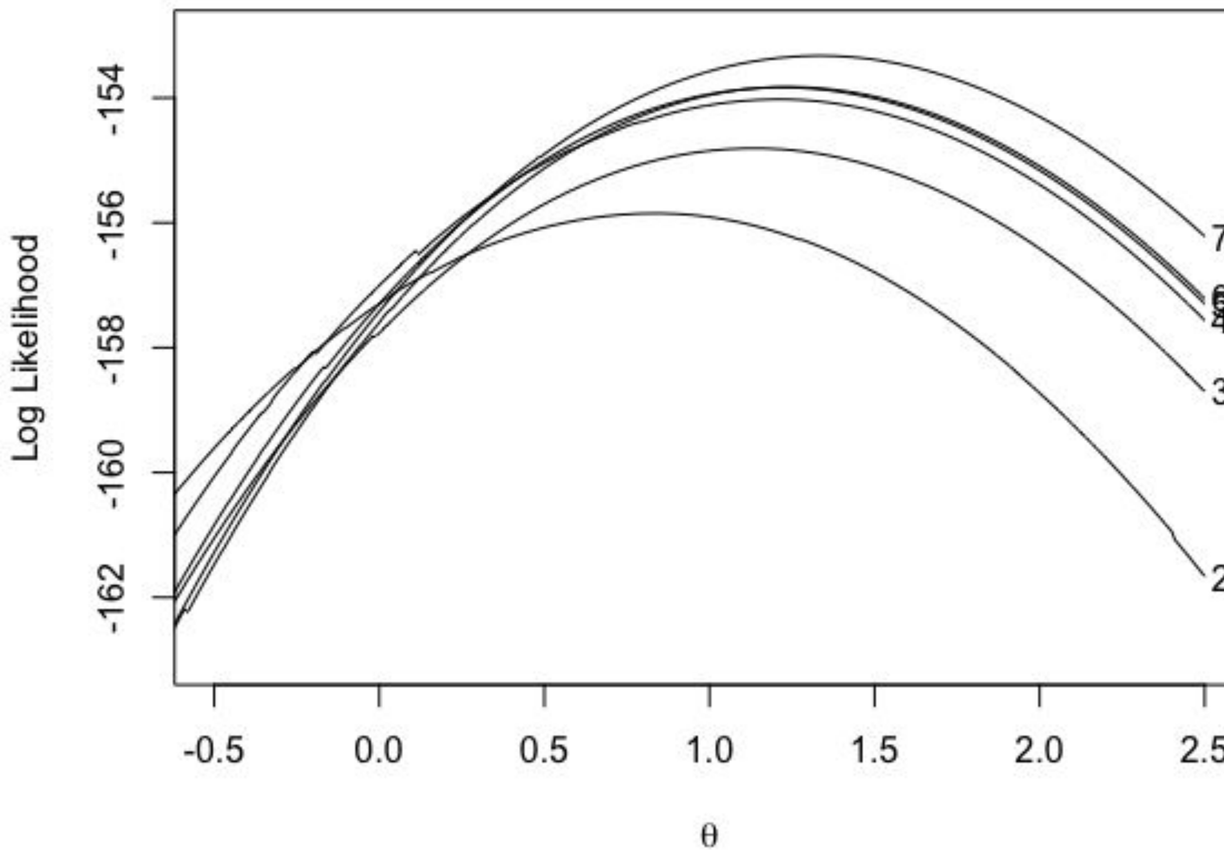
Semiparametric Models

Possibly use **Method of Sieves**
(Grenander 1981, Geman and Hwang 1982)

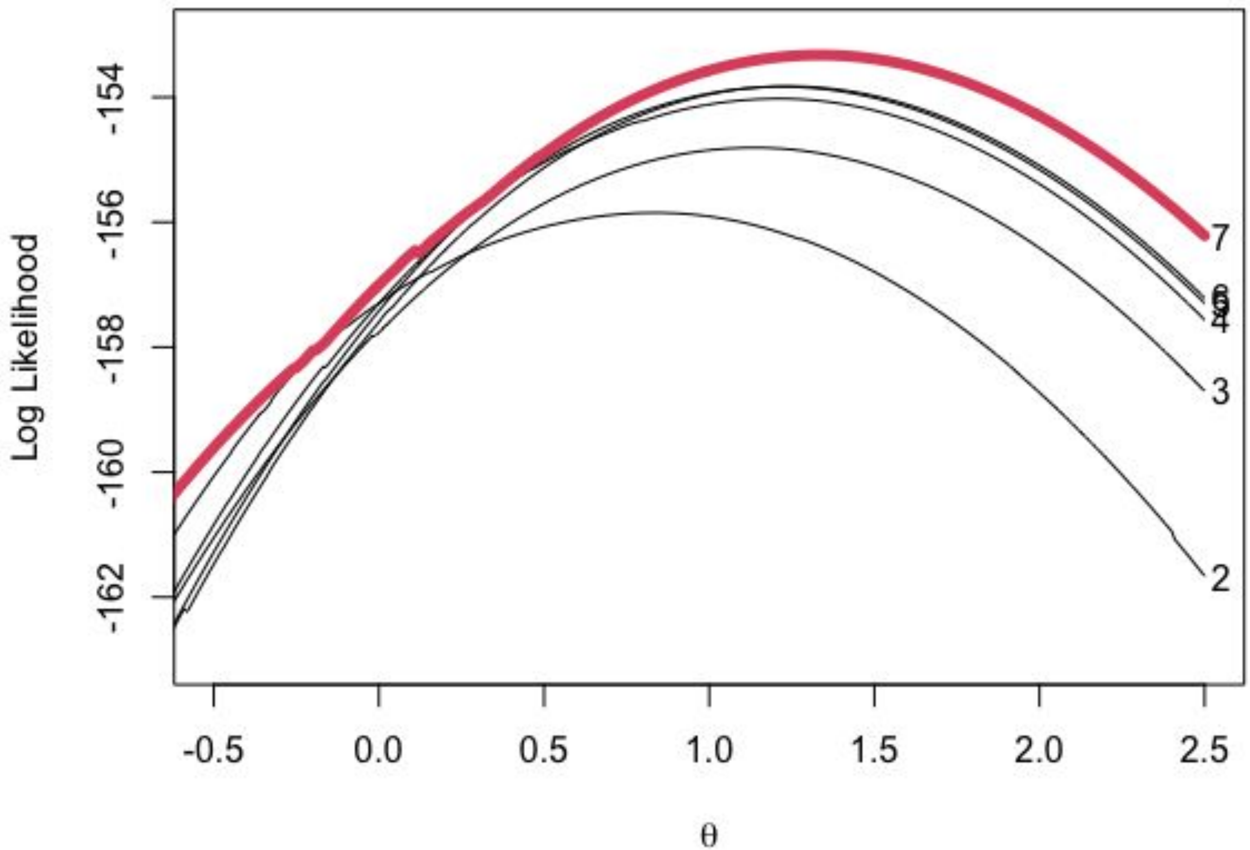
Construct **series** of $\mathcal{N}_n$ which grow in complexity with n, but slowly enough to ensure **consistency** of estimating θ

Interesting parallel with **Discrete Profiling**
(Dauncey, et al. 2015)

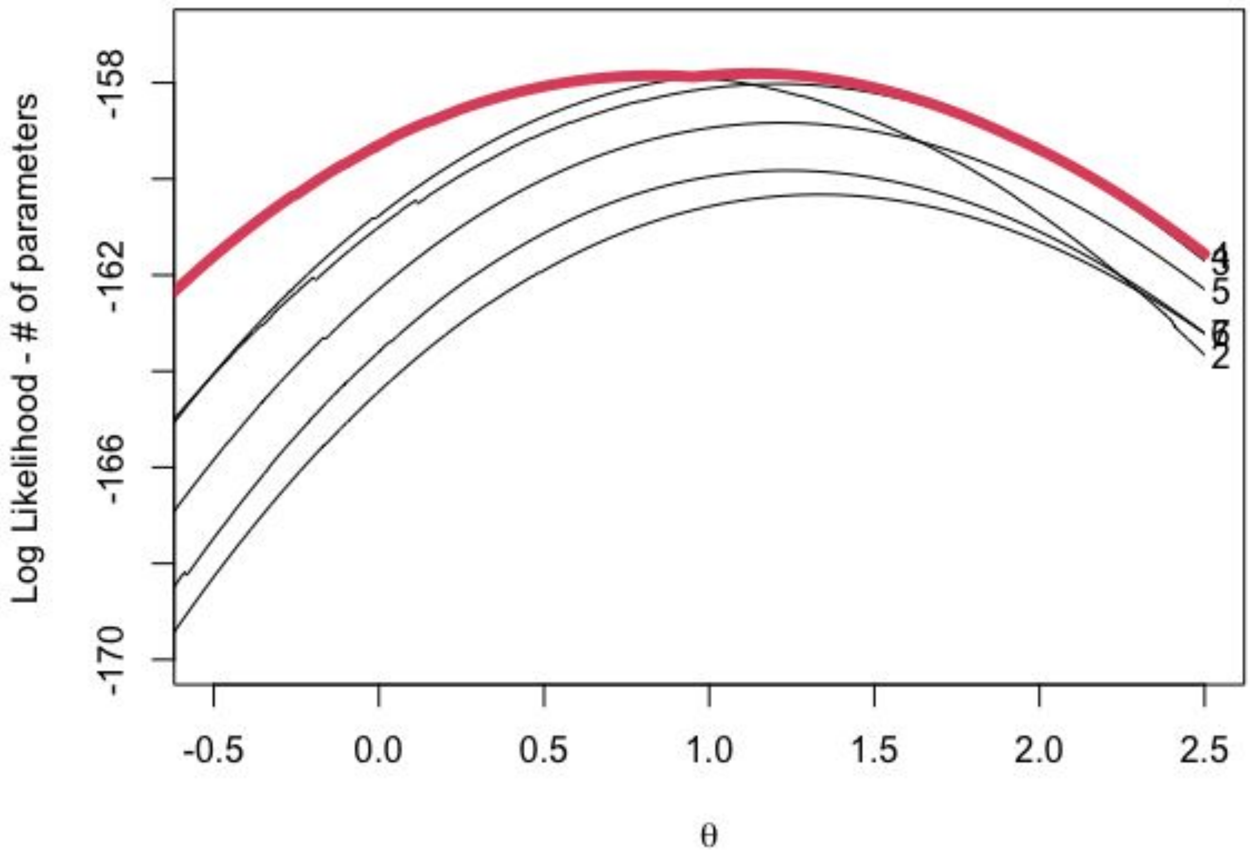
Discrete Profiling



Discrete Profiling



Discrete Profiling



Model Averaging

Another class of ideas: **Model Averaging**

Overview in Chapter 7 of **Claeskens and Hjort (2008)**

Also, **Burnham and Anderson (2002)**

Instead of fixing on one model,
average over multiple candidates

Model Averaging

Buckland, et al. (1997):

General form for **Information criterion**:

$$I_k = -2\log(L_k) + q_k$$

where q_k is a penalty term.

Then, define **model weights**:

$$w_i \propto \exp(-I_i/2) = L_i \exp(-q_i/2)$$

Model Averaging

General form for **Information criterion**:

$$I_k = -2\log(L_k) + q_k$$

where q_k is a penalty term.

For example, if

$$q_k = \log(n)p_k$$

then using **BIC**, and (see Schwarz 1978):

$$\frac{w_i}{w_j} \approx \text{Bayes factor between models } i \text{ and } j$$

Model Averaging

General form for **Information criterion**:

$$I_k = -2\log(L_k) + q_k$$

where q_k is a penalty term.

For example, if

$$q_k = 2p_k$$

then using **AIC**, and construct **Akaike weights**.

Model Averaging

Model	MLE of θ	Log likelihood at MLE	AIC	Akaike Weight
Linear	0.13	-174.79	355.58	≈ 0
Quadratic	0.84	-155.85	319.70	0.29
Cubic	1.14	-154.81	319.62	0.30
Quartic	1.21	-154.02	320.05	0.24
Quintic	1.18	-153.84	321.67	0.11
Sextic	1.24	-153.83	323.65	0.040
Septic	1.25	-153.34	324.69	0.024

Model Averaging

The **weighted estimator** for the parameter:

$$\widehat{\theta}_a = \sum_{i=1}^{K} w_i \widehat{\theta}_i$$

with variance (Equation 4.9 in Burnham and Anderson):

$$\mathrm{Var}\left(\widehat{\theta}_a\right) = \left[\sum_{i=1}^{K} w_i \sqrt{\mathrm{Var}\left(\widehat{\theta}_i | \text{model i}\right) + \left(\widehat{\theta}_i - \widehat{\theta}_a\right)^2}\right]^2$$

Model Averaging

Model	MLE of θ	Log likelihood at MLE	AIC	Akaike Weight
Linear	0.13	-174.79	355.58	≈ 0
Quadratic	0.84	-155.85	319.70	0.29
Cubic	1.14	-154.81	319.62	0.30
Quartic	1.21	-154.02	320.05	0.24
Quintic	1.18	-153.84	321.67	0.11
Sextic	1.24	-153.83	323.65	0.040
Septic	1.25	-153.34	324.69	0.024

$$1.08 \pm 1.96 \times 0.050 = (0.09, 2.07)$$

Model Averaging

Improved versions of confidence intervals in

Section 4.3.3 in Burnham and Anderson

Chapter 7 of Claeskens and Hjort

References

Buckland, et al. (1997). “Model selection: an integral part of inference. *Biometrics*. Vol. 53

Burnham and Anderson (2002). *Model Selection and Multimodel Inference.* Second Edition. Springer.

Claeskens and Hjort (2008). *Model Selection and Averaging.* Cambridge University Press.

Dauncey, et al. (2015). “Handling uncertainties in background shapes: the discrete profiling method.” *Journal of Instrumentation*, Vol. 10, Issue 4.

Geman and Hwang (1982). "Nonparametric Maximum Likelihood Estimation by the Method of Sieves." *The Annals of Statistics*, Vol. 10, No. 2, pp. 401-414.

Grenander (1981). *Abstract Inference*. Wiley.

Murphy and van der Vaart (2000). “On Profile Likelihood.” *Journal of the American Statistical Association*. Vol. 95

Schwarz (1978). “Estimating the dimension of a model.” *The Annals of Statistics*. Vol. 6