The Loss Rank Principle for Model Selection

Marcus Hutter

Canberra, ACT, 0200, Australia

http://www.hutter1.net/

ANU

RSISE

NICTA

Contents

- Model Complexity Selection
- Empirical Model Selection
- Theoretical Model Selection
- The Loss Rank Principle for Model Selection
- LoRP for Classification and (Linear) Regression
- Experimental Results
- Outlook and References

Abstract

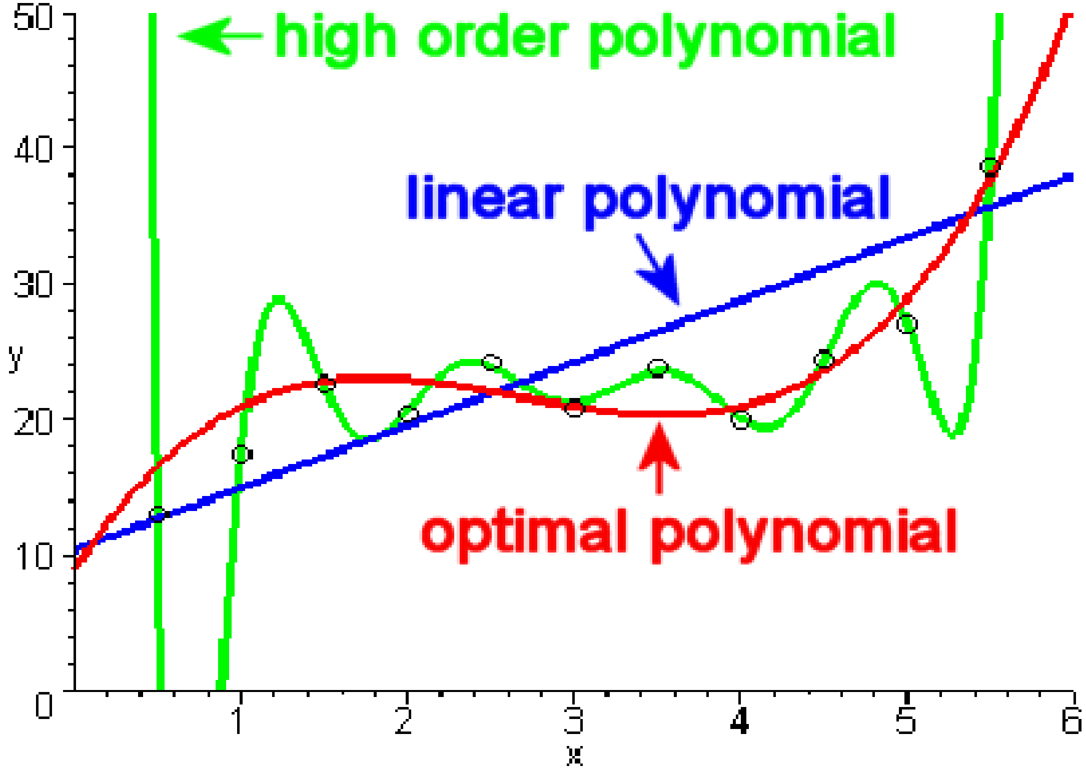
A key issue in statistics and machine learning is to automatically select the "right" model complexity, e.g. the number of neighbors to be averaged over in k nearest neighbor (kNN) regression or the polynomial degree in regression with polynomials. We suggest a novel principle (LoRP) for model selection in regression and classification. It is based on the loss rank, which counts how many other (fictitious) data would be fitted better. LoRP selects the model that has minimal loss rank. Unlike most penalized maximum likelihood variants (AIC,BIC,MDL), LoRP only depends on the regression functions and the loss function. It works without a stochastic noise model, and is directly applicable to any non-parametric regressor, like kNN.

Example: Polynomial Regression

- Straight line does not fit data well (large training error)
  high bias ⇒ poor predictive performance

- High order polynomial fits data perfectly (zero training error)
  high variance (overfitting)
  ⇒ poor prediction too!

- Reasonable polynomial degree d performs well.
  How to select d?
  minimizing training error obviously does not work.



Model Complexity Selection

Regression and classification:
Learn functional relation $f:\mathcal{X}\to\mathcal{Y}$ for data $D=\{(x_1,y_1)...(x_n,y_n)\}$.

Model complexity: Many regression models are controlled by some smoothness or flexibility or complexity parameter c:

- Non-parametric example: the number of neighbors to be averaged over in k nearest neighbor (kNN) regression.
- Parametric example: the polynomial degree d in regression with polynomials.

Model selection: Select the “right” model complexity c, like k or d.

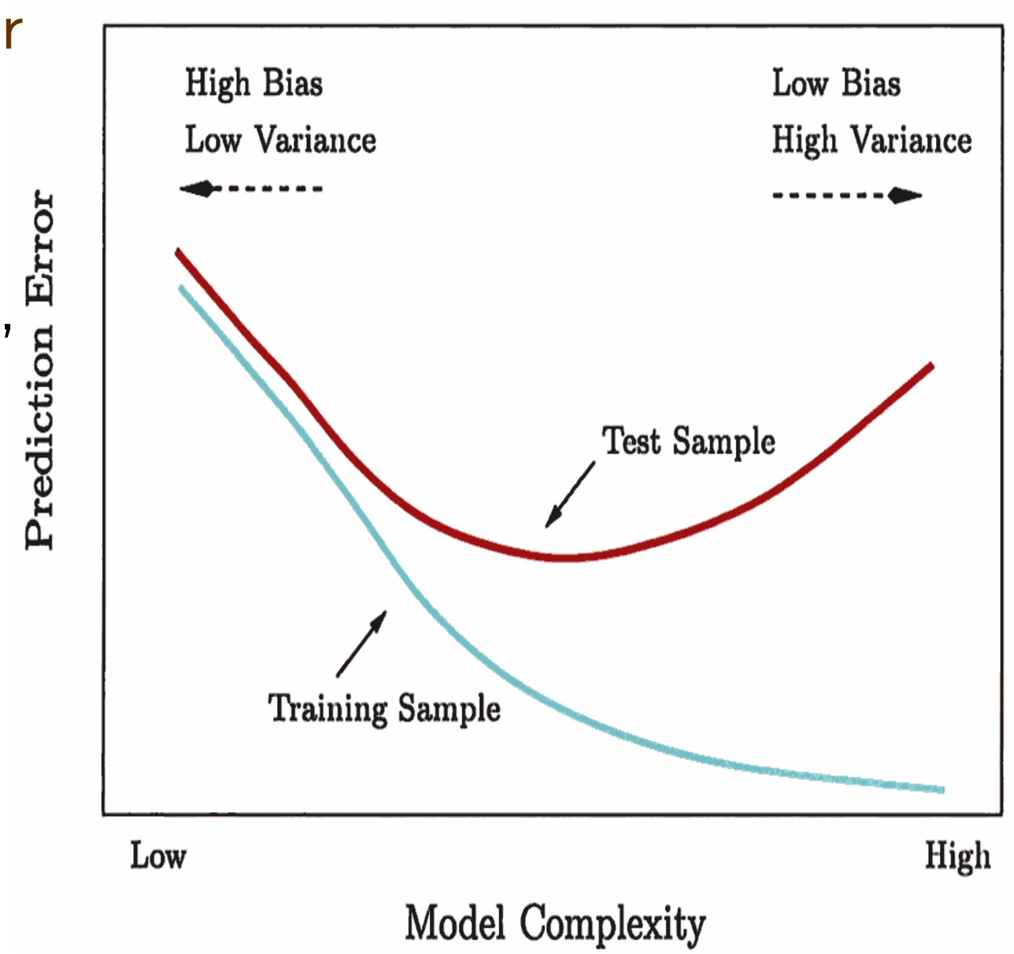
Selection cannot be based on the training error,
since the more complex the model (large d, small k)
the better the fit on D (perfect for $d=n$ and $k=1$).

This problem is called overfitting,
for which various remedies have been suggested ...

Empirical Model Selection

How to select complexity parameter

- Kernel width a,
- penalization constant λ,
- number k of nearest neighbors,
- the polynomial degree d?



Empirical test-set-based methods:

Regress on training set and minimize empirical error w.r.t. "complexity" parameter (a, λ, k, d) on a separate test-set.

Problems: Reduces training set size and therefore regression quality.

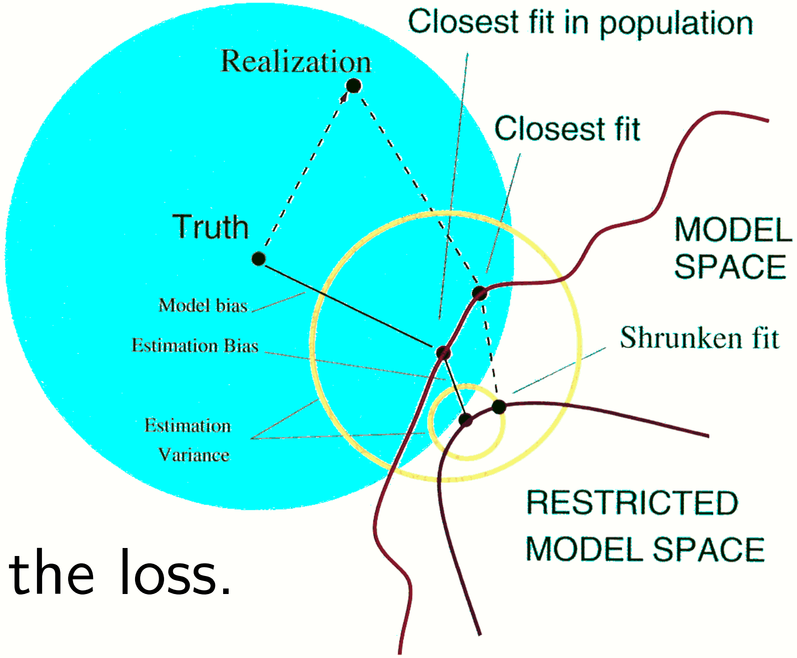
Theoretical Model Selection

How to select complexity or flexibility or smoothness parameter:

Kernel width a, penalization constant λ, number k of nearest neighbors, the polynomial degree d?

For parametric regression with d parameters

- Bayesian model selection,
- Akaike Information Criterion (AIC),
- Bayesian Information Criterion (BIC),
- Minimum Description Length (MDL),



They all add a penalty proportional to d to the loss.

Problems:

- Limited to (semi)parametric models (with d “true” parameters).
- Needs a full stochastic model $\mathrm{P}(\boldsymbol{y}|\text{parameters})$, not just f.
- Loss function is often not exploited.

The Loss Rank Principle for Model Selection

Let $\hat{f}_D^c : \mathcal{X} \to \mathcal{Y}$ be the (best) regressor of complexity c on data D.

The loss *Rank* of $\hat{f}_D^c$ is defined as the number of other (fictitious) data D' that are fitted better by $\hat{f}_{D'}^c$ than D is fitted by $\hat{f}_D^c$.

- c is too small $\Rightarrow$ $\hat{f}_D^c$ fits D badly $\Rightarrow$ many other D' can be fitted better $\Rightarrow$ *Rank* is large.
- c is too large $\Rightarrow$ many D' can be fitted well $\Rightarrow$ *Rank* is large.
- c is appropriate $\Rightarrow$ $\hat{f}_D^c$ fits D well *and* not too many other D' can be fitted well $\Rightarrow$ *Rank* is small.

LoRP: Select model complexity c that has minimal loss *Rank*

Unlike most penalized maximum likelihood variants (AIC,BIC,MDL),

- LoRP only depends on the regression and the loss function.
- It works without a stochastic noise model, and
- is directly applicable to any non-parametric regressor, like kNN.

LoRP for Classification (finite $\mathcal{Y}$)

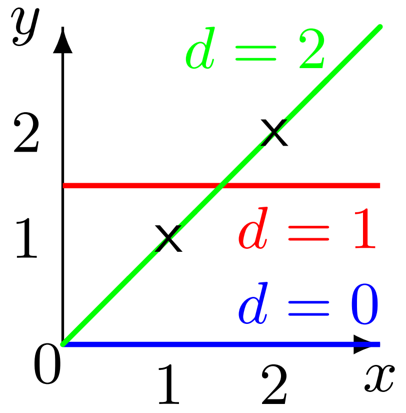
- Observed data: $D = (\boldsymbol{x},\boldsymbol{y}) := ((x_1,y_1)...(x_n,y_n)) \in (\mathcal{X}\times\mathcal{Y})^n =: \mathcal{D}$
- $y_i \approx f_{true}(x_i)$ are distorted from the unknown true values $f_{true}(x_i)$.
- Regression: Gives $r:\mathcal{D}\to\mathcal{F}$ such that $\hat{y} := r(x|D) \equiv \hat{f}_D(x) \approx f_{true}(x)$ for *all* $x\in\mathcal{X}$.
- Empirical loss of regressor r: $L := \text{Loss}_r(\boldsymbol{y}|\boldsymbol{x}) \;:=\; \text{Loss}(\boldsymbol{y},\hat{\boldsymbol{y}}) \;=\; \sum_{i=1}^n \text{Loss}(y_i, r(x_i|\boldsymbol{x},\boldsymbol{y}))$
- Loss Rank: $\text{Rank}_r(\boldsymbol{y}|\boldsymbol{x}) := \#\{\boldsymbol{y}'\in\mathcal{Y}^n : \text{Loss}_r(\boldsymbol{y}'|\boldsymbol{x}) \leq L\}$
- Class of regressors $\mathcal{R}\ni r$, e.g. kNN $\{r_k : k\in I\!N\}$, or $\{r_d =$best poly. of degree $d : d\in I\!N_0\}$.

$$\boxed{\textbf{LoRP: } r^{best} \;=\; \arg\min_{r\in\mathcal{R}} \text{Rank}_r(\boldsymbol{y}|\boldsymbol{x})}$$

Example: Simple Discrete

Consider $\mathcal{X} = \{1,2\}$, $\mathcal{Y} = \{0,1,2\}$, $x_1 = y_1 = 1$, $x_2 = y_2 = 2$, $n = 2$, least squares (zero, constant, linear) polynomials $\mathcal{R} = \{r_d : d = 0,1,2\}$, and quadratic Loss:

d	$r_d(x\mid\boldsymbol{x},\boldsymbol{y}')$	$\text{Loss}_d(\boldsymbol{y}'\mid\boldsymbol{x})$	$\text{Loss}_d(D)$
0	0	${y'_1}^2 + {y'_2}^2$	5
1	$\frac{1}{2}(y'_1 + y'_2)$	$\frac{1}{2}(y'_2 - y'_1)^2$	$\frac{1}{2}$
2	$(y'_2 - y'_1)(x-1) + y'_1$	0	0



	$\text{Rank}_{r_d}(y'_1y'_2\mid 12)$									
d	1	2	3	4	5	6	7	8	9	$\text{Rank}_{r_d}(D)$
0	$y'_1y'_2 = 00$	< 01	$= 10$	< 11	< 02	$= 20$	< 21	$= \mathbf{12}$	< 22	8
1	$y'_1y'_2 = 00$	$= 11$	$= 22$	< 01	$= 10$	$= 21$	$= \mathbf{12}$	< 02	$= 20$	7
2	$y'_1y'_2 = 00$	$= 01$	$= 02$	$= 10$	$= 11$	$= 20$	$= 21$	$= 22$	$= \mathbf{12}$	9

So LoRP selects r_1 as best regressor, since it has minimal rank on D. r_0 fits D too badly and r_2 is too flexible (perfectly fits all D').

LoRP for Regression

continuous / measure space $\mathcal{Y}$ (mostly $\mathcal{Y}=I\!\!R$)

Define $\mathrm{Rank}_r(\boldsymbol{y}|\boldsymbol{x}) := |V_r(L)| :=$ Volume$(\{\boldsymbol{y}'\in\mathcal{Y}^n : \mathrm{Loss}_r(\boldsymbol{y}'|\boldsymbol{x})\leq L\})$.

Problem: Rank is often infinite.

Solution:

- Regularization by adding $\alpha||\boldsymbol{y}||^2$ to the Loss.
- Determine α by minimizing Rank_r^α w.r.t. α.

Example: Simple Continuous

Consider the Simple Discrete Example but with interval $\mathcal{Y}=[0,2]$.

The first table (r_d and Loss_d) remains unchanged,
while the second table becomes:

d	$V_d(L)=\{\boldsymbol{y}'\in[0,2]^2:...\}$	$\lvert V_d(L)\rvert$	$\lvert V_d(\mathrm{Loss}_d(D))\rvert$
0	${y'_1}^2+{y'_2}^2\leq L$	$2\sqrt{\max\{L-4,0\}}+$ $L(\frac{\pi}{4}-\cos^{-1}(\min\{\frac{2}{\sqrt{L}},1\}))$	$\doteq 3.6$
1	$\frac{1}{2}(y'_2-y'_1)^2\leq L$	$4\sqrt{2L}-2L$	3
2	$0\leq L$	4	4

So LoRP again selects r_1 as best regressor, since it has smallest loss volume on D.

LoRP for Linear Models

- Examples: kNN, Kernel, polynomial, linear basis function (LBF) regression are all linear in $\boldsymbol{y}$, i.e. $\hat{\boldsymbol{y}} = M\boldsymbol{y}$ for some matrix $M(\boldsymbol{x})$.
- Rank is volume of ellipsoid $\{\hat{\boldsymbol{y}}^\top(\mathbb{1}-M)^\top(\mathbb{1}-M)\hat{\boldsymbol{y}}\} \leq L$.

$\Rightarrow$ efficient $O(n^3)$ algorithm for $\det(\mathbb{1}-M)$.

- LoRP for projective regression: $O(n)$ algorithm.
- For Gaussian LBF reg, LoRP is similar to Bayesian model selection.
- Other non-parametric model selection schemes [MacKay92,HTF01]: Use $\mathrm{tr}M$ as effective degrees of freedom d in AIC/BIC/MDL.
- Problem: $d_{eff} = \mathrm{tr}M$ is a heuristic that breaks down already for simple examples like “k nearest neighbors *excluding* the closest neighbor” ($\mathrm{tr}M \equiv 0$).

LoRP for Variable Selection

- Variable=feature=attribute selection $\mathcal{S} \subseteq \{\infty, ..., \lceil\}$:
  Linear regression model $y = \sum_{j\in\mathcal{S}} \beta_j x_j + noise$.

- Model consistency:
  LoRP with optimized α for $n \to \infty$ selects the the true model.

- Asymptotic mean efficiency: For suitable choice of α, LoRP estimates the regression function asymptotically (mean) efficiently.

Variable Selection for Simulated Data

Percentage of correctly-fitted linear models over 1000 replications.

LoRP compared to Akaike and Bayesian Information Criterion (AIC&BIC)

n	d	SNR	AIC	BIC	**LoRP**	n	d	SNR	AIC	BIC	**LoRP**
100	5	1	62	62	69	300	5	1	74	82	83
		5	85	85	86			5	78	90	91
		10	80	90	91			10	81	94	94
	10	1	52	42	54		10	1	63	67	71
		5	63	77	77			5	70	85	86
		10	68	84	85			10	74	90	90
	20	1	32	22	36		20	1	54	45	61
		5	55	63	65			5	64	79	80
		10	56	73	74			10	67	85	85

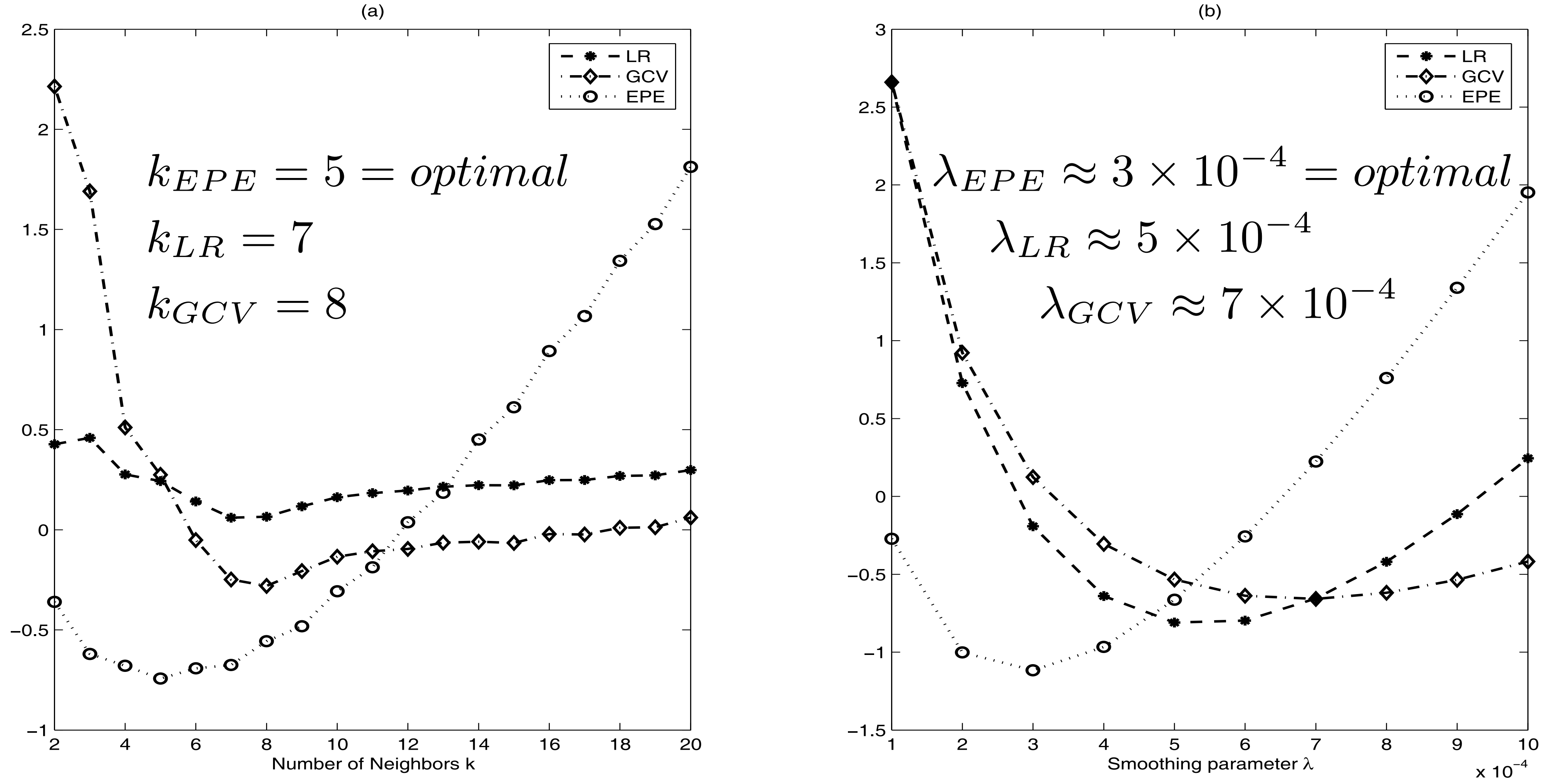
Fourier Order Selection for Simulated Data

Est. of mean efficiency over 1000 replications of $y = \log(1-x)+noise$.

LoRP compared to Akaike and Bayesian Information Criterion (AIC&BIC)

n	σ	AIC	BIC	**LoRP**	n	σ	AIC	BIC	**LoRP**
400	.001	1.00	.98	.99	600	.001	1.00	.98	1.00
	.01	.93	.68	.90		.01	.99	.67	.92
	.05	.88	.67	.95		.05	.90	.66	.94
	.1	.88	.67	.92		.1	.90	.67	.93
	.5	.81	.66	.85		.5	.82	.66	.83
	1	.79	.63	.82		1	.79	.65	.82
	5	.67	.65	.70		5	.65	.67	.66
	10	.54	.67	.59		10	.54	.59	.54
	100	.31	.89	.33		100	.40	.90	.41

Nearest Neighbor and Spline Regression



$k_{EPE} = 5 = optimal$

$k_{LR} = 7$

$k_{GCV} = 8$

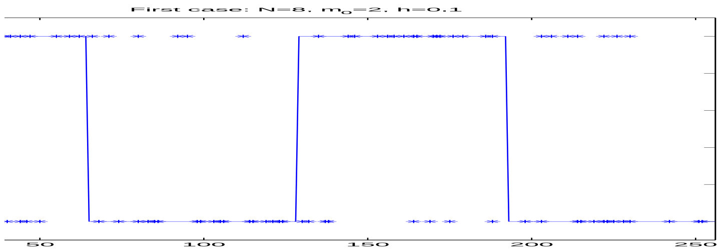
$\lambda_{EPE} \approx 3 \times 10^{-4} = optimal$

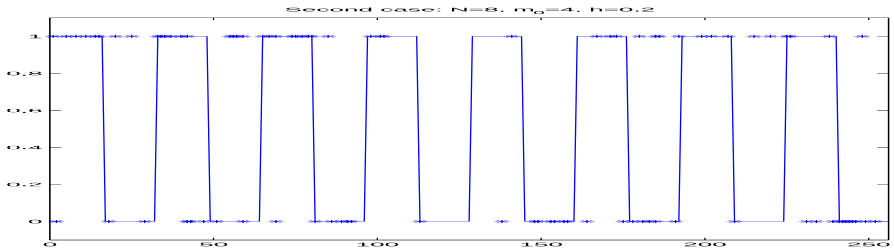
$\lambda_{LR} \approx 5 \times 10^{-4}$

$\lambda_{GCV} \approx 7 \times 10^{-4}$

- EPE=Expected Prediction Error = infeasible gold standard
- GCV=Generalized Cross Validation, • LR=Loss Rank,
- $y = f(x) = \sin(12(x+0.2))/(x+0.2) + noise, \quad x \in [0;1]$.

LoRP for Classification

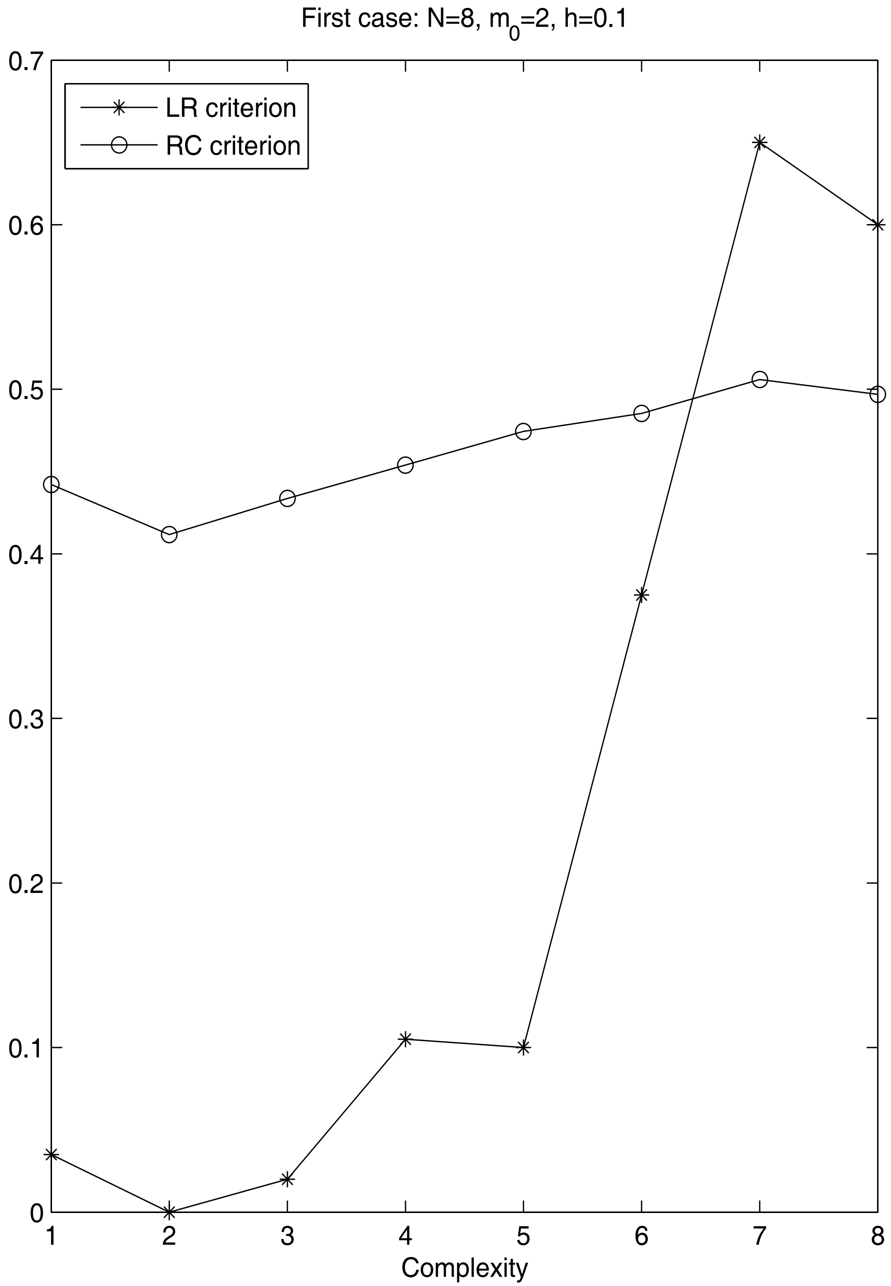




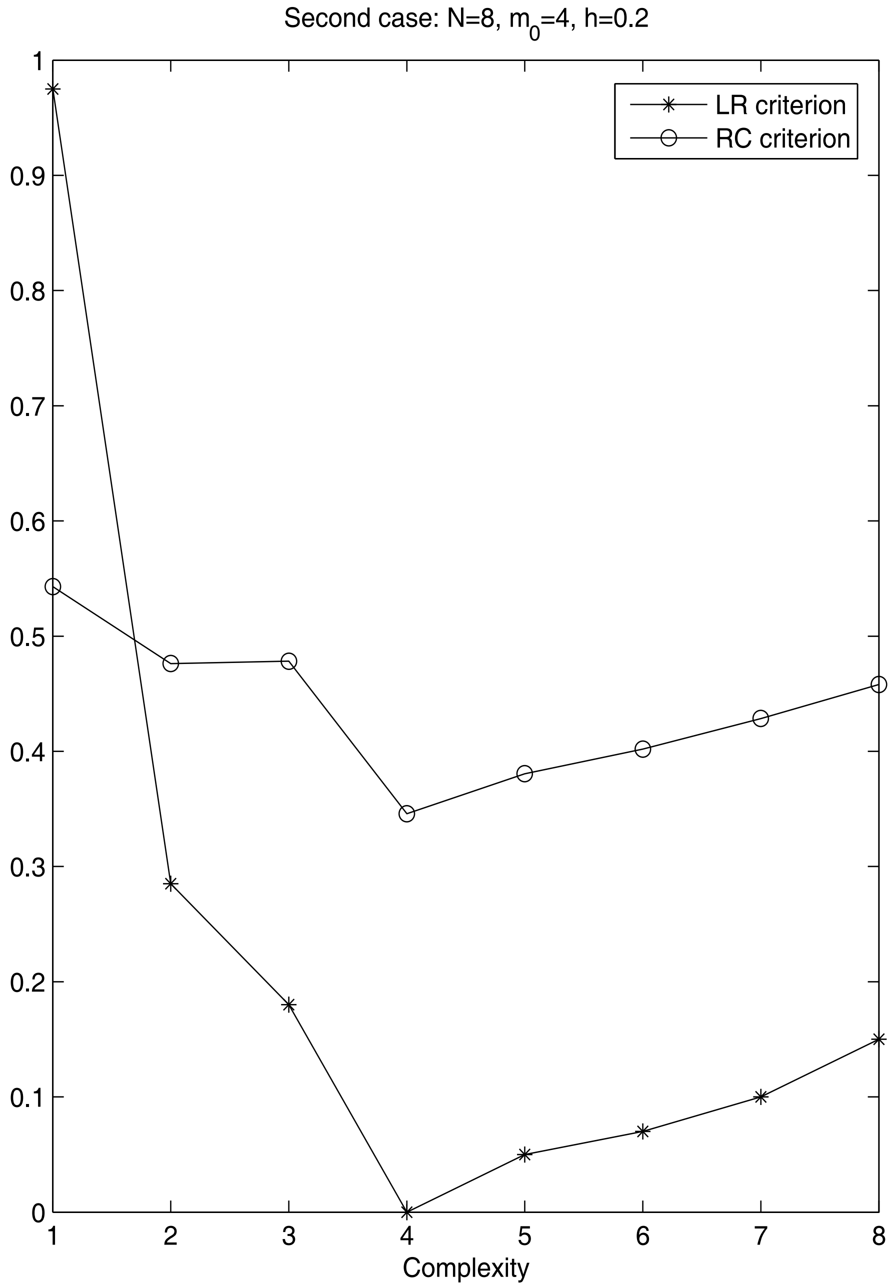
Proportions of correct identification of LR and Rademacher Criterion

n	h	**LR**	RC	n	h	**LR**	RC
50	.05	.12	.13	200	.05	.23	.21
	.1	.35	.35		.1	.67	.66
	.2	.62	.64		.2	.99	.97
	.3	.95	.97		.3	1	1
100	.05	.15	.15	300	.05	.30	.28
	.1	.41	.41		.1	.78	.76
	.2	.89	.90		.2	1	.99
	.3	.98	.98		.3	1	1

LoRP for Classification



First case: N=8, m_0=2, h=0.1



Second case: N=8, m_0=4, h=0.2

LoRP for Clustering

- A natural loss is the within-cluster sum of dissimilarities.
- Comparison of LR=Loss Rank criterion to CH= Calinski and Harabasz (1974) criterion on
- Simulated data with 2–4 Gaussian clusters in 2D of varying σ.
- Using Monte Carlo simulation to compute the Loss Rank.

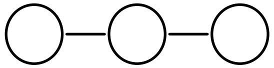
Percentages of correct identification over 100 replications

#Cl	σ	CH	**LR**	#Cl	σ	CH	**LR**	#Cl	σ	CH	**LR**
2	1	1	0.82	3	1	0.99	0.84	4	1	0.92	0.56
	2	1	0.74		2	0.7	0.45		2	0.04	0.38
	3	1	0.86		3	0	0.39		3	0	0.50

LoRP for Graphical Modeling

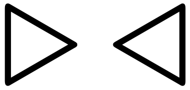
Graphical models ≈ Markov networks
≈ graphical log-linear modeling ≈ Bayesian networks

Goal: learn/recover correct graphical structure.

Example: 100 samples from graph 

Proportions of correct identification:

n	200	500	1000	2000	5000
LR	.2	.7	.9	1	1
BIC	.05	.4	.7	.8	1

Similar results for $n = 10000$ and 6 vertex graph 

Compute Bootstrap Loss Rank by Monte Carlo sampling

Outlook

LoRP seems to be a promising principle with a lot of potential

- Preliminary experiments look promising.
- Rank for non-linear LoRP can be estimated by Monte Carlo.
- $\det(1\!\!1 - M)$ may be approximated numerically in time $O(n)$.
- Explicit expressions for kNN on a grid.
- LoRP for hybrid model classes.
- Cubic algorithm for linear LoRP works for general additive loss.
- LoRP can be used to select the loss-function itself.
- (Log) Rank can be regarded as a code (length) of $\boldsymbol{y}$.

Other ideas that count: normalized ML, luckiness framework, empirical Rademacher complexity, permutation tests, Bootstrapping.

Thanks! Questions? Details:

[Hut07] M. Hutter. The loss rank principle for model selection. In Proc. 20th Annual Conf. on Learning Theory (COLT'07), volume 4539 of LNAI, pages 589–603, San Diego, 2007. Springer, Berlin.

[HT10] M. Hutter and M. Tran. Model selection with the loss rank principle. Computational Statistics and Data Analysis, 54:1288–1306, 2010.

[TH10] M. Tran and M. Hutter. Model selection by loss rank for classification and unsupervised learning. arXiv, 1011(1379):1–20, 2010.