CONVEX RISK MINIMIZATION METHODS AND ROBUSTNESS

ANDREAS CHRISTMANN

University of Dortmund, Department of Statistics, Germany

Christmann@statistik.uni-dortmund.de

DoMuS Workshop, Dortmund, 18/NOV/2003

Forschungsband Modellbildung und Simulation

UNIVERSITÄT DORTMUND SFB 475 Komplexitätsreduktion in multivariaten Datenstrukturen

Joint work with:

Ingo Steinwart (Los Alamos National Laboratory, USA)

CONTENTS

ZDF: 6. November 2003

Optimieren der KFZ-Versicherung

. . . Seit über 20 Jahren sind die Tarifmerkmale in der Fahrzeugversicherung (Kasko) nahezu unverändert. Jetzt wird der Kaskoschutz 'runderneuert'. Wichtigste Änderung: Die Typklassen-Struktur wird mit statistischen Verfahren neu ermittelt und berücksichtigt wie die KFZ-Haftpflichtversicherung nun auch Merkmale wie zum Beispiel Fahrleistung und Garage. . . .

(Thomas J. Kramer)

1. EXAMPLE: MOTOR VEHICLE INSURANCE

- Project in SFB-475 (with Dr. A. Kovac, University of Essen)
- Verband öffentlicher Versicherer, Düsseldorf, Germany
  non-aggregated data from 15 insurance companies:
  3 GB, > 6 millions obs., > 70 explanatory variables, many discrete
- What is the expected claim amount? $\Rightarrow$ insurance tariffs
- What is the probability of a claim?
- complex dependencies, empty cells, missing values
- some extreme high costs

STATISTICAL OBJECTIVES

- Y claim size [year], x vector of explanatory variables
- Actual premium charged to the customer:
  pure premium + safety loading + administrative costs + desired profit
- Primary response: Pure premium. $E(Y|X = x)$
- Secondary response: Prob. of claim. $P(Y > 0|X = x)$

Exploratory Data Analysis

Cost per policy holder per year: total mean ≈ 360 EUR

Cost	% obs.	% of total sum	Mean	Median	Std Dev
total			364	0	21996
0	94.9	0	0	0	0
(0,2000]	2.2	6.7	1097	1110	520
(2000,10000]	2.4	27.1	4156	3496	1940
(10000,50000]	0.4	19.8	18443	15059	8911
>50000	0.07	46.4	234621	96417	784365

Maximum cost: $>$ 27 Million EUR !

Define $C \in \{0, \ldots, k+1\}$ for 'no claim', 'small claim', ..., 'extreme claim'

$$E(Y|X=x) = P(C=k+1) \cdot E(Y|X=x, C=k+1) + P(C \neq k+1) \cdot \sum_{c=1}^{k} P(C=c|X=x) \cdot E(Y|X=x, C=c)$$

2. CONVEX RISK MINIMIZATION BASED ON KERNELS

Vapnik '98

data set $(x_i, y_i) \in \mathbb{R}^p \times \{-1, +1\}$, assume (X_i, Y_i) i.i.d. $\mathbb{P}$, $\mathbb{P}$ unknown,

predictor $\hat{f}(x)$, classifier $\operatorname{sign}(\hat{f}(x) + \hat{b})$, loss function $L(y, f(x) + b)$

Goal: $\arg\min_f \mathrm{E}_{\mathbb{P}} L(Y, f(X) + b)$

1^{st} idea: $\arg\min_f \frac{1}{n} \sum_{i=1}^n I(y_i, f(x_i) + b)$

but: $I(y, f + b)$ not convex, problem often NP-hard (Höffgen et al. '95)

2^{st} idea: minimize regularized empirical risk

$$(\hat{f}_{n,\lambda}, \hat{b}_{n,\lambda}) = \arg\min_{f \in H,\, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i) + b) + \lambda \|f\|_H^2,$$

where L convex, $\lambda > 0$ regularization parameter,

H reproducing kernel Hilbert space (RKHS) of kernel k

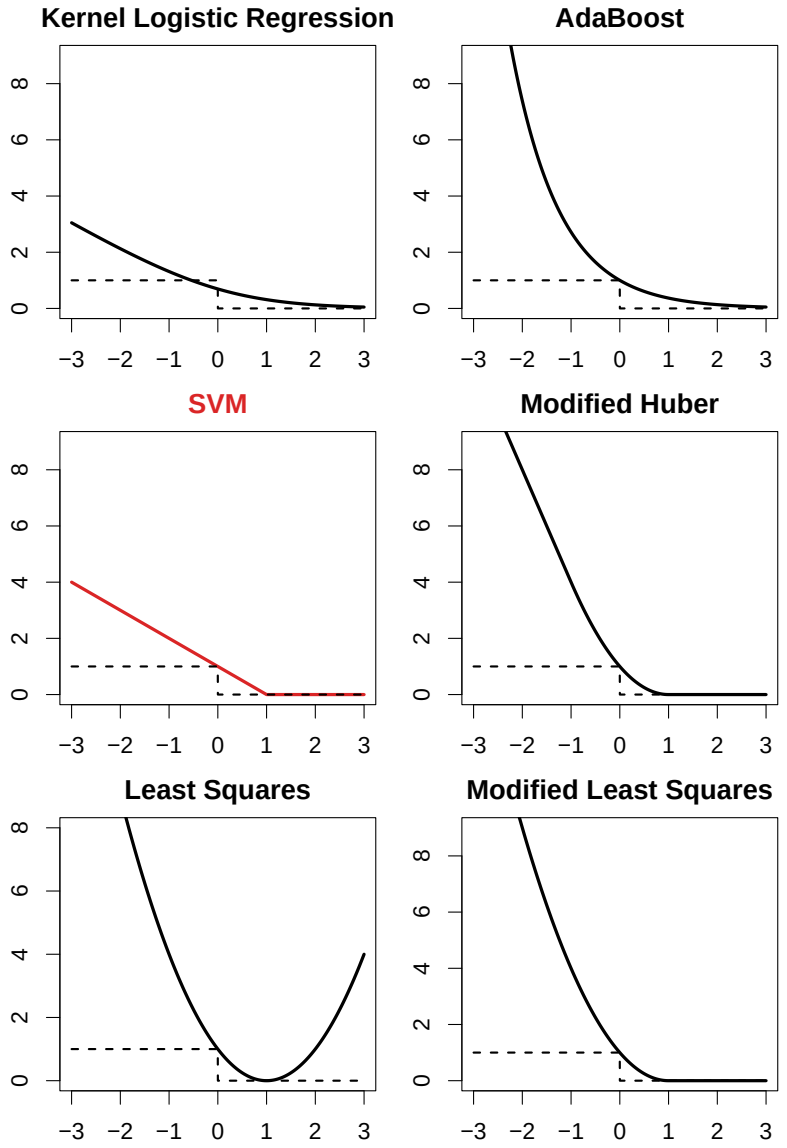
reg. theor. $(f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda}) = \arg\min_{f \in H,\, b \in \mathbb{R}} \mathrm{E}_{\mathbb{P}} L(Y, f(X) + b) + \lambda \|f\|_H^2$

risk: Vapnik '98, Zhang '01, Steinwart '02: universal consistency

Special loss functions:

Method	L, $v = y(f(x) + b)$
Kernel Logistic Regression	$\ln(1 + \exp(-v))$
AdaBoost	$\exp(-v)$
Support Vector Machine	$\max(1 - v, 0)$
Modified Huber	$-4v$, if $v < -1$
	$\max(1 - v, 0)^2$, else
Least Squares	$(1 - v)^2$
Modified Least Squares	$\max(1 - v, 0)^2$

Vapnik '98,
Schölkopf & Smola '02,
Freund & Schapire '96,
Friedman, Hastie & Tibshirani '00,
Hastie, Tibshirani & Friedman '01,
Suykens et al. '02,
Zhang '01, ...



SUPPORT VECTOR MACHINE (SVM) (Vapnik '98)

Dual program is convex & quadratic !

$$\begin{aligned} \arg\min \quad & \tfrac{1}{2}\alpha' Q \alpha - \alpha' \mathbf{1} \\ \text{s.t.:} \quad & \tfrac{1}{n}\textstyle\sum_i \alpha_i y_i = 0 \\ & \alpha_i \in [0, C] \\ & \text{where } (Q)_{ij} = y_i y_j \, k(x_i, x_j),\ Q \in \mathbb{R}^{n \times n} \text{ !} \end{aligned}$$

linear kernel: $k(x_i, x_j) = x_i' x_j$

RBF kernel: $k(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2)$, $\gamma > 0$, $\quad u := ||x_i - x_j||$

Software:

many computational tricks: e.g. Sequential Minimization Optimization (SMO)

Overview: `http://www.kernel-machines.org`

SVM^{light}: Joachims '99, mySVM: Rüping '00 (Computer Science, Univ. of Dortmund)

R: e1071

$\varepsilon-$SV REGRESSION (Vapnik '98)

$$\frac{1}{2}\left|\left|\theta\right|\right|^2 + C\frac{1}{n}\sum_i L_\varepsilon(x_i, y_i, f) = \min!$$

$\varepsilon-$insensitive loss: $L_\varepsilon(x, y, f) = \max\left\{0, |y - f(x)| - \varepsilon\right\}$

$\nu-$SV REGRESSION (Schölkopf et al. '00)

$$\frac{1}{2}\left|\left|\theta\right|\right|^2 + C\left(\nu\varepsilon + \frac{1}{n}\sum_i L_\varepsilon(x_i, y_i, f)\right) = \min!$$

KERNEL LOGISTIC REGRESSION (KLR)

- SVM estimates: $\operatorname{sign}\left(P(Y=1|X=x)-\frac{1}{2}\right)$
- KLR estimates: $f(x)=\log\left(\frac{P(Y=1|X=x)}{P(Y=-1|X=x)}\right)$
- KLR vs. SVM:
  $\longrightarrow$ classification performance similar to SVM
  $\longrightarrow$ offers estimate of class probabilities: risk scoring
  $\longrightarrow$ computationally more expensive.
  Keerthi et al. '02: fast dual algorithm with pseudo-code
  myKLR: Rüping '03 (Computer Science, Univ. of Dortmund)
  `http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYKLR/`
  $n=10^5$ manageable on PC
  $\longrightarrow$ in general: no. SV's $\approx$ no. obs. for KLR fit $\hat{f}(x)=\hat{b}+\sum_i \hat{\alpha}_i k(x,x_i)$,
  i.e. no data compression. For SVM: in general no. SV's $\ll$ no. obs.

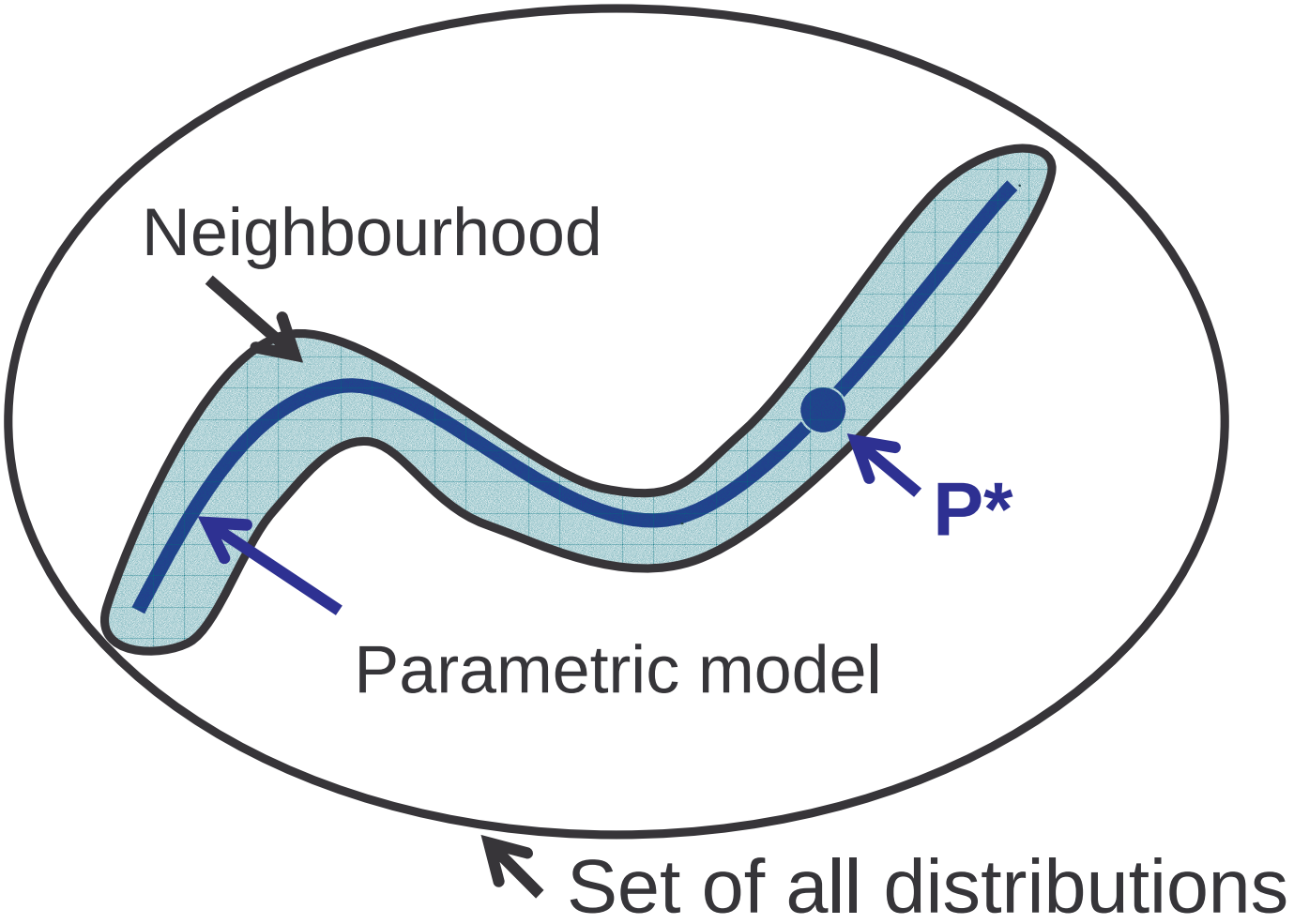
3. ROBUSTNESS ASPECTS

"Points that are not support vectors have no influence, so that in non-degenerate cases slight pertubations of such points will not affect the solution."

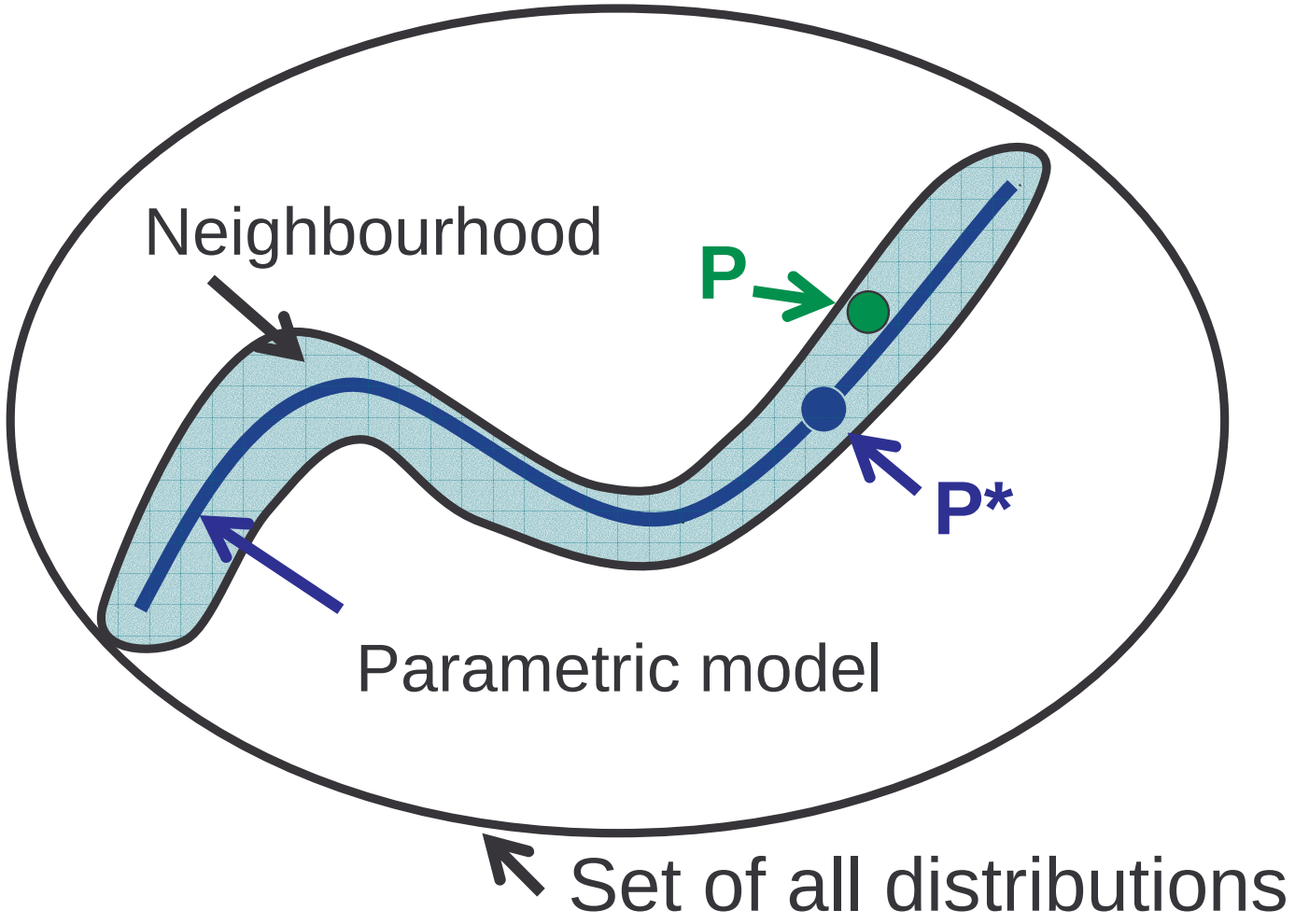
Cristianini & Shawe-Taylor (2000), p. 97.

PROP. 1: Suppose θ can be expressed in terms of the SVs which are not at bound, i.e. $\theta = \sum_{i=1}^{n} \gamma_i x_i$, with $\gamma_i \neq 0$, only if $\alpha_i \in (0, \frac{1}{n})$ (where α_i from solution of dual problem). Then local movements of any margin error x_m parallel to θ do *not* change the hyperplane.

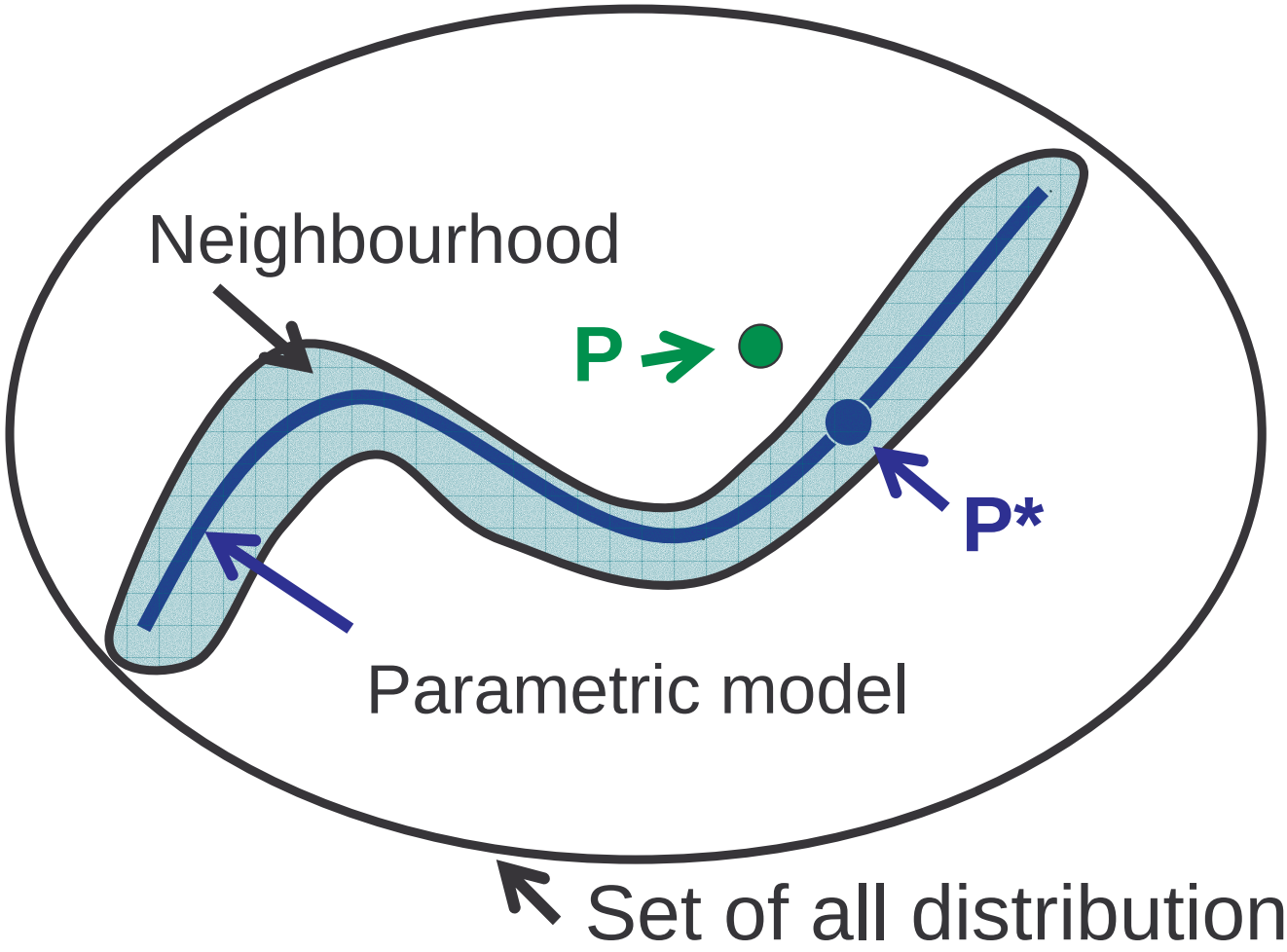
Proof: e.g. Schölkopf & Smola '02
Note: pertubation of the point carried out in feature space. Hence depends of the specific kernel.



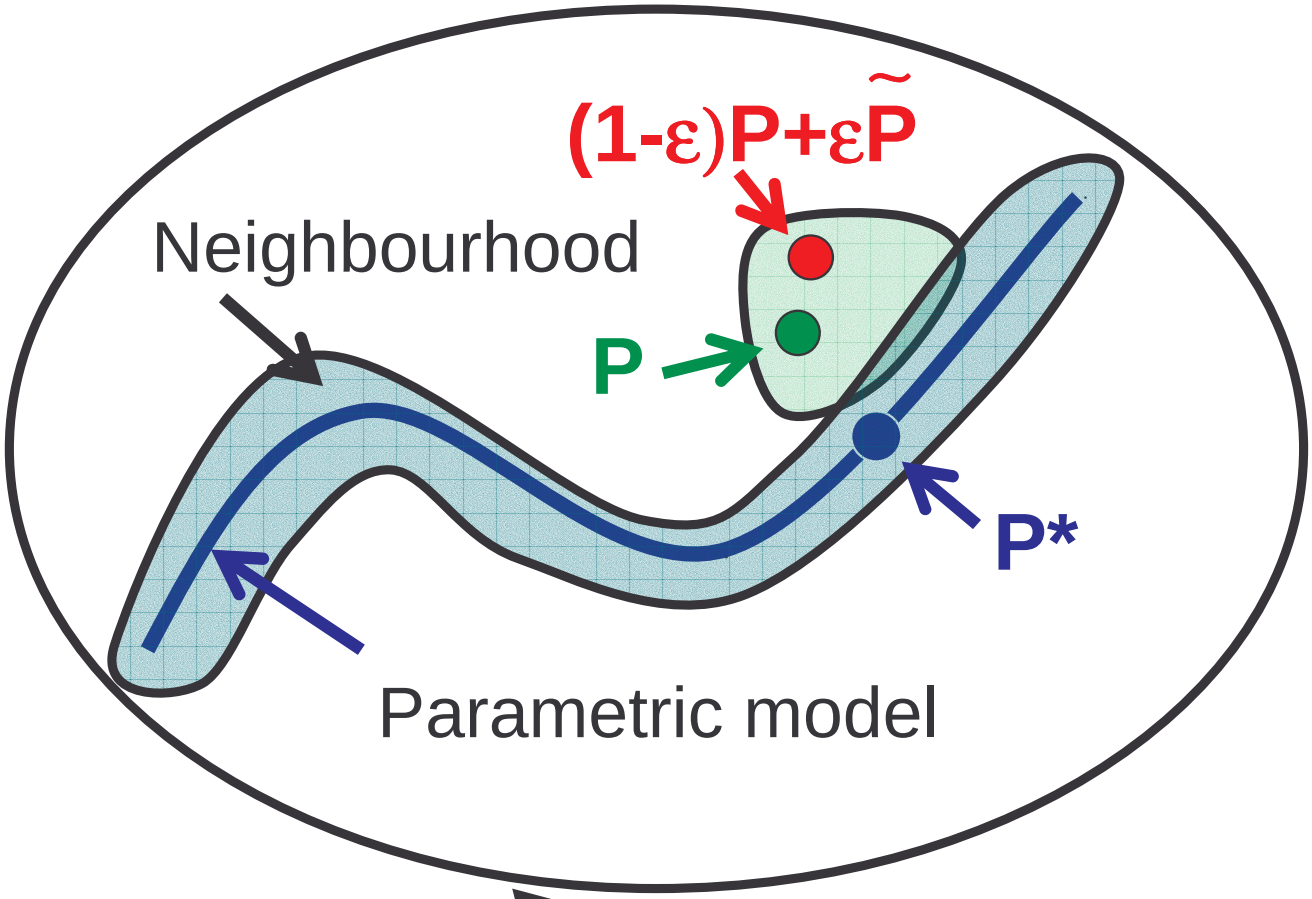
Goal: Statistic T(P)



Goal: Statistic T(P)



Goal: Statistic T(P)



Goal: Statistic T(P)

Hampel's influence function:

$$IF(z; T, \mathbb{P}) = \lim_{\varepsilon \downarrow 0} \frac{T\big((1-\varepsilon)\mathbb{P} + \varepsilon \Delta_z\big) - T(\mathbb{P})}{\varepsilon}$$

Here: $T(\mathbb{P})$ is regularized theoretical risk:

$$(f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda}) = \arg \min_{f \in H,\, b \in \mathbb{R}} \mathrm{E}_{\mathbb{P}}\, L(Y, f(X) + b) + \lambda \| f \|_H^2$$

Tukey's sensitivity curve:

$$SC_n(z) = n\, [T_n(z_1, \ldots, z_{n-1}, z) - T_{n-1}(z_1, \ldots, z_{n-1})]$$

PATTERN RECOGNITION

PROP. 2: Existence of influence function.
Assume: $L : Y \times \mathbb{R} \to [0, \infty)$ convex loss function with $L'' > 0$ and continuous, $X \subset \mathbb{R}^p$ compact, H is RKHS of continuous kernel.
Then the influence function of the classifiers minimizing the theoretical regularized risk exists for all $z \in X \times Y$.

Proof: Chr & Steinwart '03.

PROP. 3: Uniform bounds on the difference quotient used by IF.
Assume: $L : Y \times \mathbb{R} \to [0, \infty)$ continuous and convex loss function, $X \subset \mathbb{R}^p$ compact, H is RKHS of continuous kernel.
Then for all $\lambda > 0$ there exists a constant $c_L(\lambda) > 0$ (explicitly known) such that for ALL distributions $\mathbb{P}$ and $\tilde{\mathbb{P}}$ on $X \times Y$ we have

$$\left\| \frac{f_{(1-\varepsilon)\mathbb{P}+\varepsilon\tilde{\mathbb{P}},\lambda} - f_{\mathbb{P},\lambda}}{\varepsilon} \right\|_H \leq c_L(\lambda) \, \|\mathbb{P} - \tilde{\mathbb{P}}\|_{\mathcal{M}}, \quad \varepsilon > 0 .$$

Proof: Chr & Steinwart '03

Applications:

- SVM, KLR, ...
- Tukey's sensitivity curve: $\mathbb{P} = \mathbb{P}_n$, $\tilde{\mathbb{P}} = \Delta_z$, $\varepsilon = \frac{1}{n}$
- upper bound of max-bias curve:

$$\left\| f_{(1-\varepsilon)\mathbb{P}+\varepsilon\tilde{\mathbb{P}},\lambda} - f_{\mathbb{P},\lambda} \right\|_H \leq \varepsilon c_L(\lambda) \, \|\mathbb{P} - \tilde{\mathbb{P}}\|_{\mathcal{M}}$$

PROP. 4: Uniform bounds for the IF.
Assume: $L : Y \times \mathbb{R} \to [0, \infty)$ convex, twice cont. diff. loss function with $0 < a \leq L'' \leq b < \infty$, $X \subset \mathbb{R}^p$ compact, H is RKHS of continuous kernel. Then for all $\lambda > 0$ there exists a constant $c_L(\lambda) > 0$ such that for ALL distributions $\mathbb{P}$ on $X \times Y$ we have

$$\|\mathrm{IF}(z; T, \mathbb{P})\|_{H \times \mathbb{R}} \leq c_L(\lambda) \, \|\mathbb{P} - \Delta_z\|_{\mathcal{M}} \, .$$

Proof: Chr & Steinwart '03

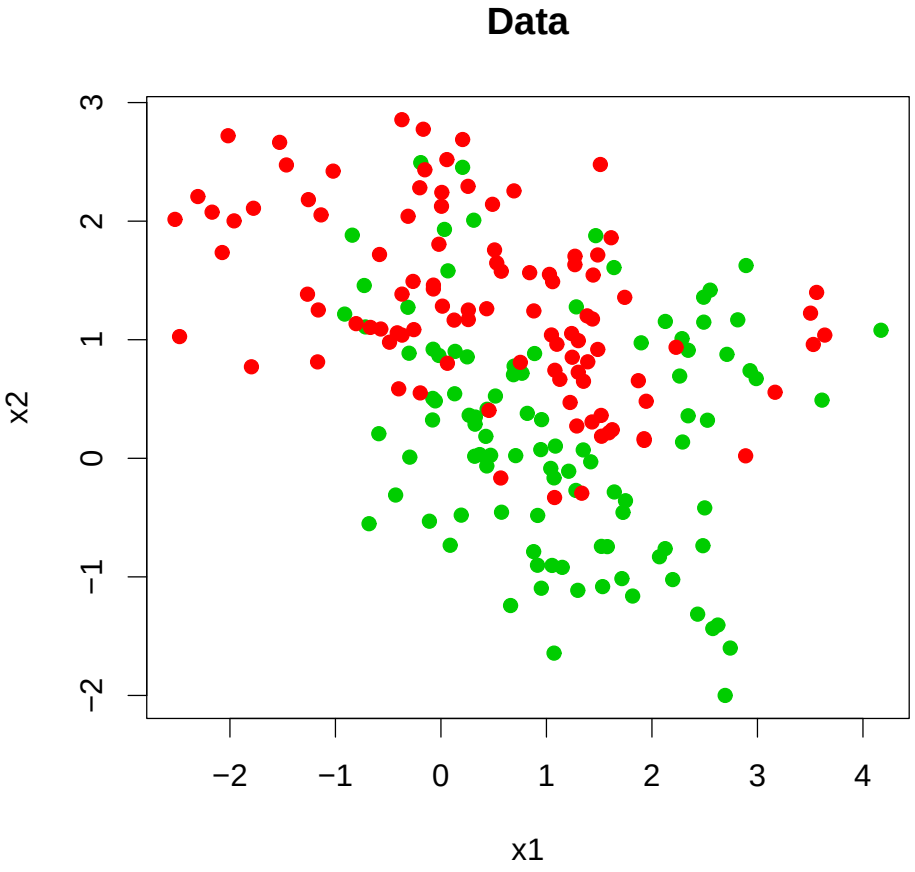
- SVM is excluded in Prop. 4. But LS-SVM is special case.
- Similar results without intercept b.
- If $\mathbb{P}$ and $\tilde{\mathbb{P}}$ have densities: last 2 results can be specified w.r.t. Hellinger metric because

$$\|\mathbb{P} - \tilde{\mathbb{P}}\|_{\mathcal{M}} \leq 2 \, H(\mathbb{P}, \tilde{\mathbb{P}}) \leq 2 \, \|\mathbb{P} - \tilde{\mathbb{P}}\|_{\mathcal{M}}^{1/2} ,$$

where $H(\mathbb{P}, \tilde{\mathbb{P}}) = [\int (\sqrt{p} - \sqrt{\tilde{p}})^2 \, d\nu]^{1/2}$.

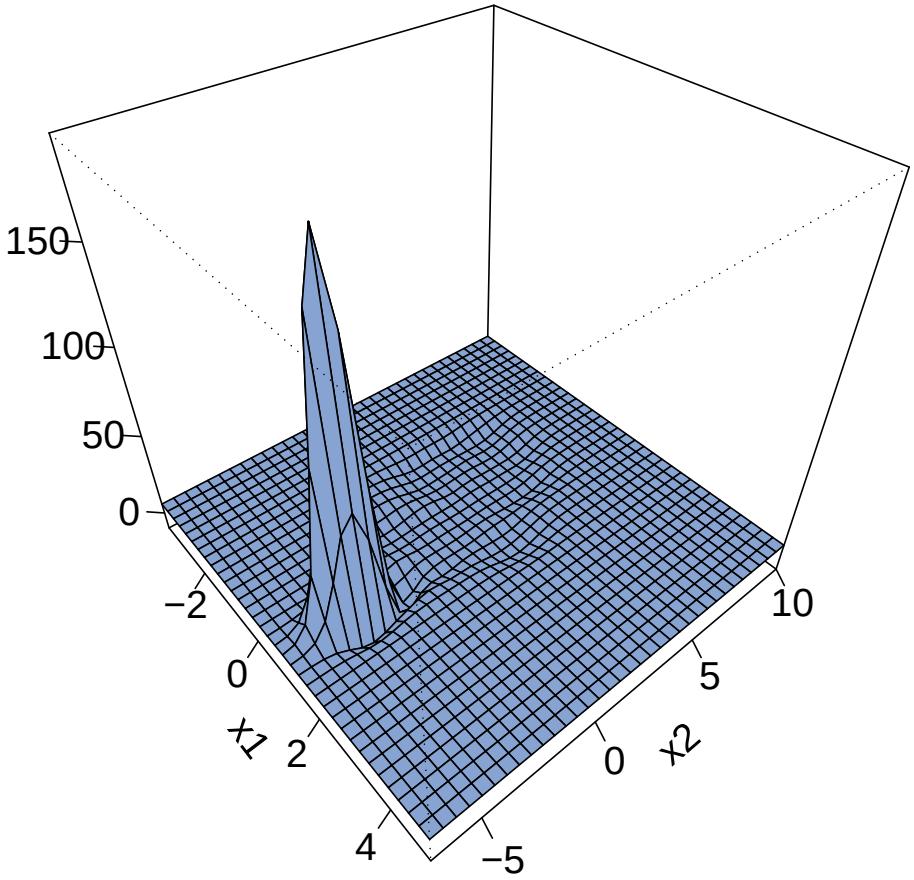
TOY EXAMPLE

Mixture data
(Hastie, Tibshirani, Friedman '01)



● $y = -1$, ● $y = +1$

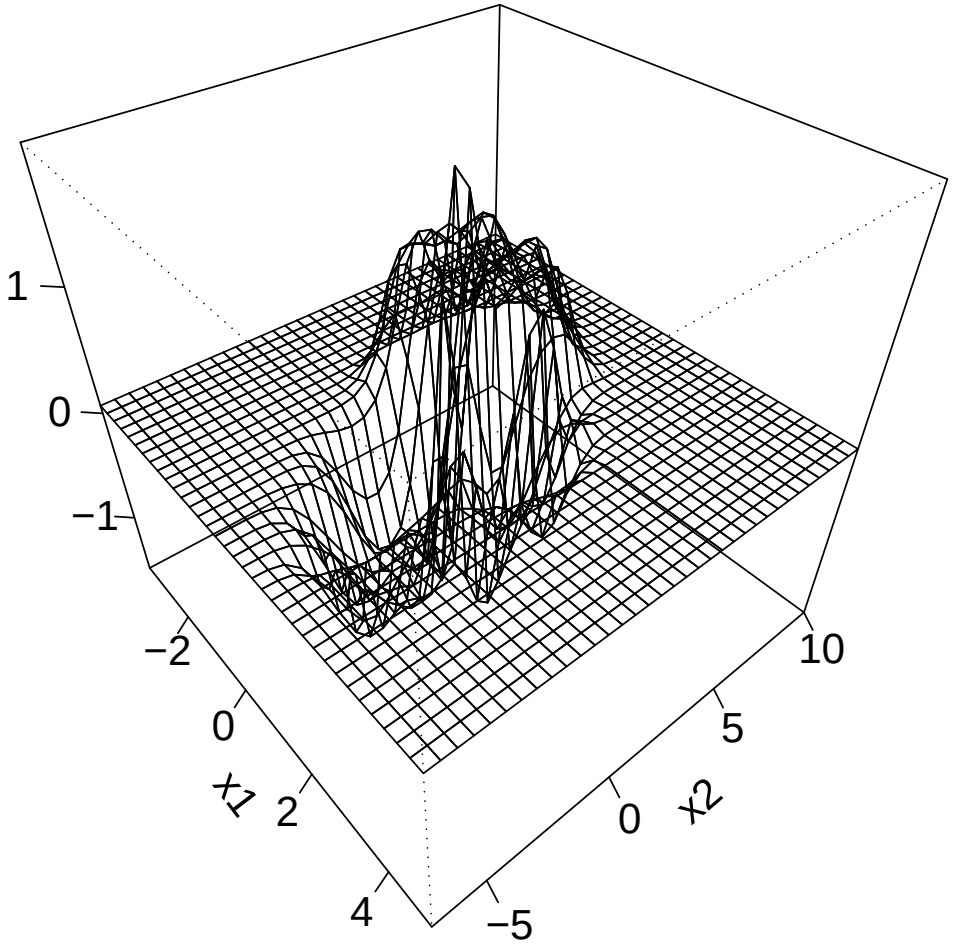
Sensitivity curve of $\hat{f}(x) + \hat{b}$ based on SVM-RBF



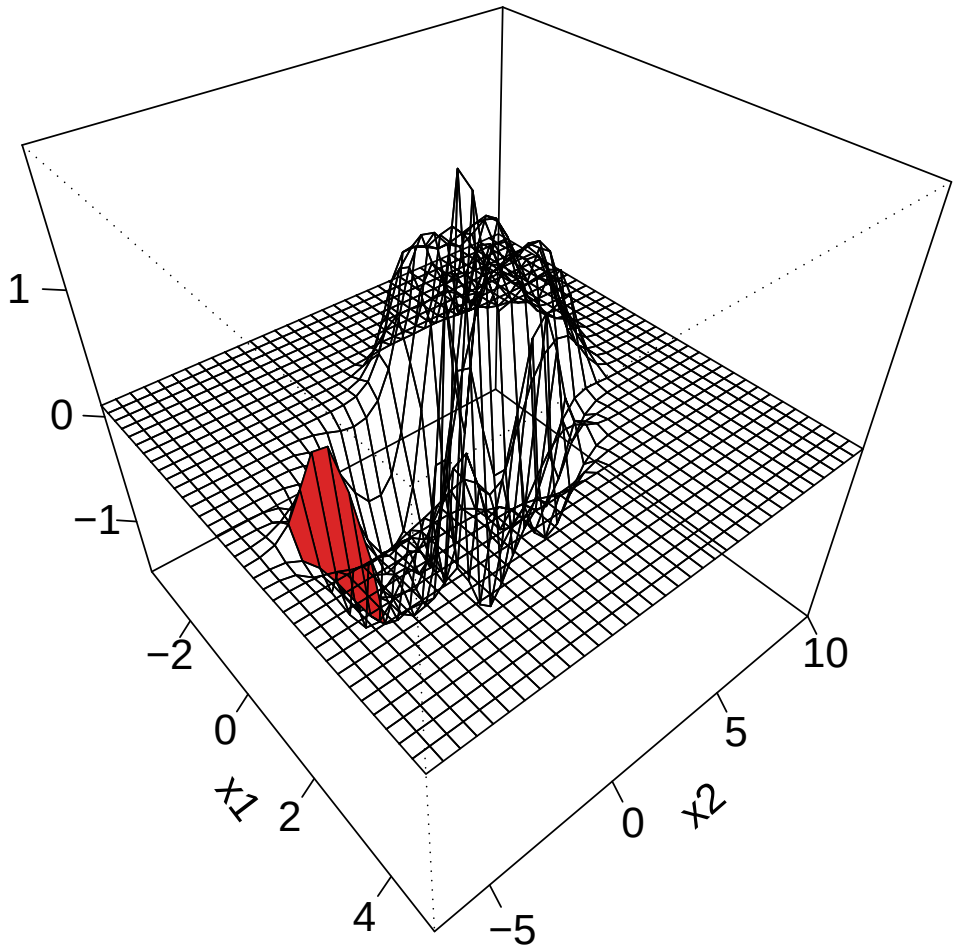
new obs at $x_1 = 2$, $x_2 = -2$, $y = +1$

$\hat{f}(x) + \hat{b}$ based on SVM-RBF

original data set

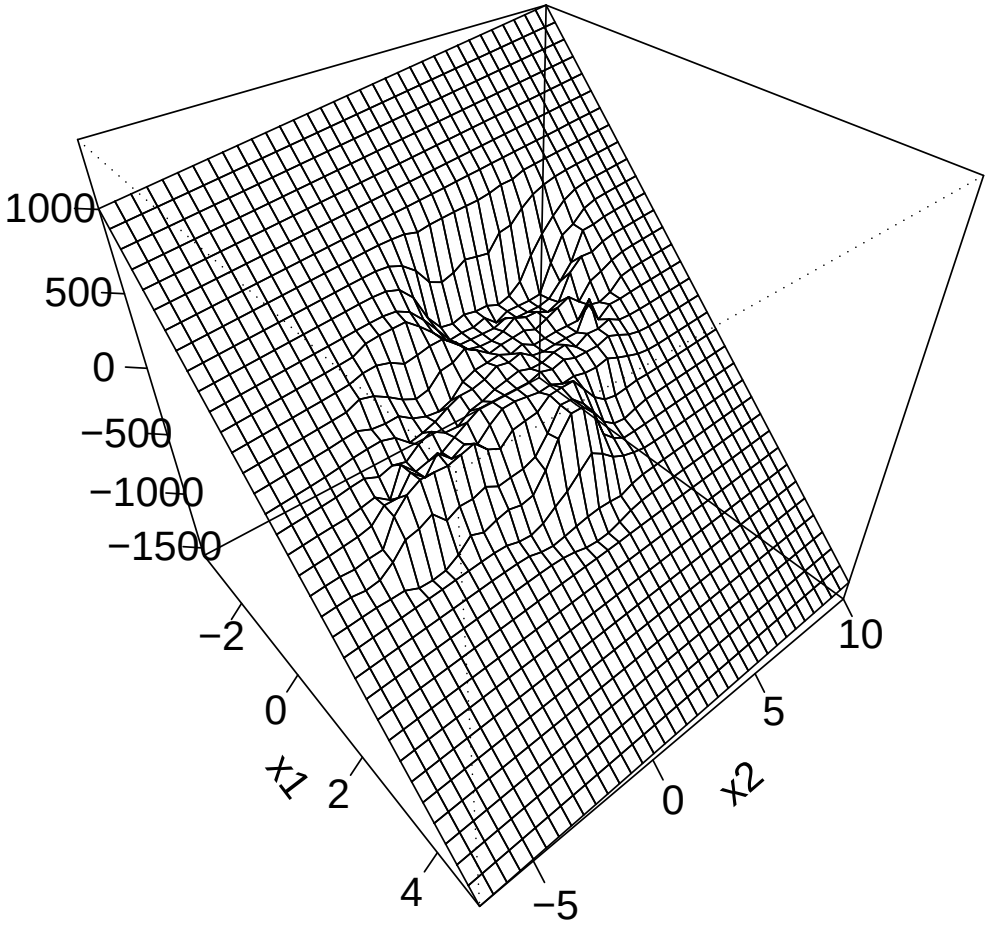


modified data set

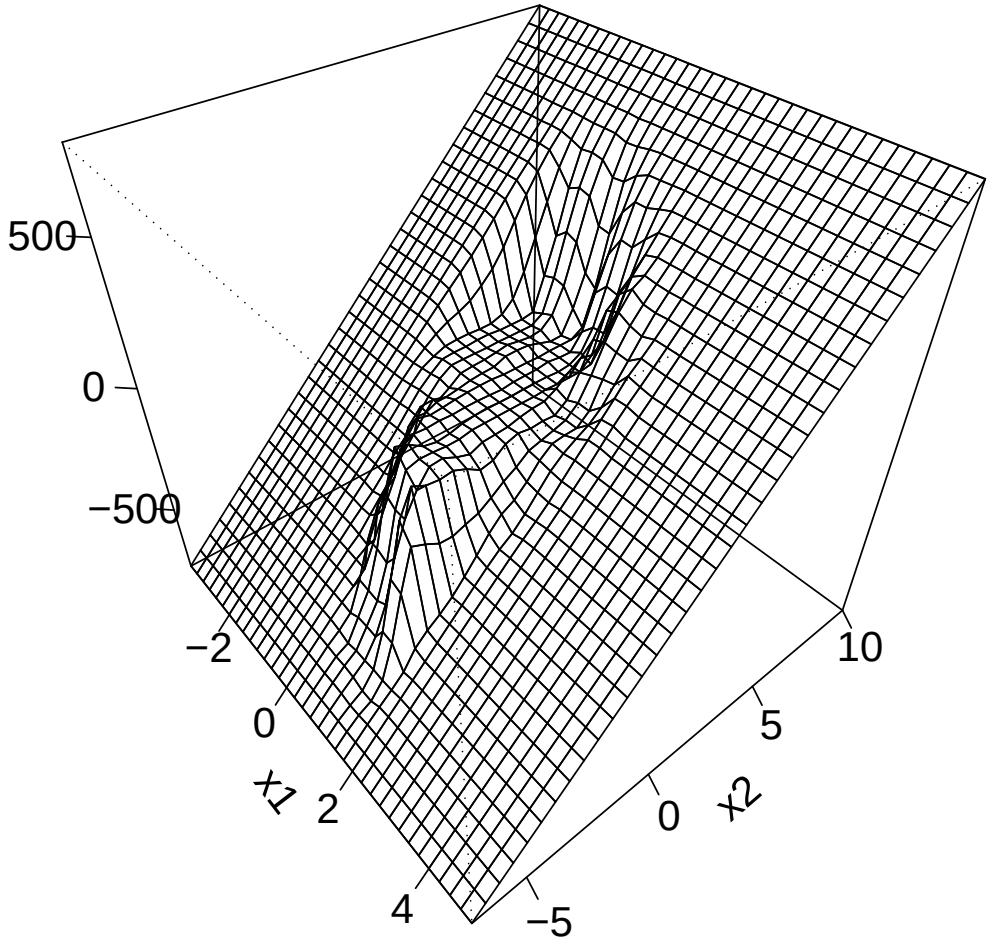


new obs at $x_1 = 2$, $x_2 = -2$, $y = +1$

Sensitivity curve for intercept $\hat{b}$



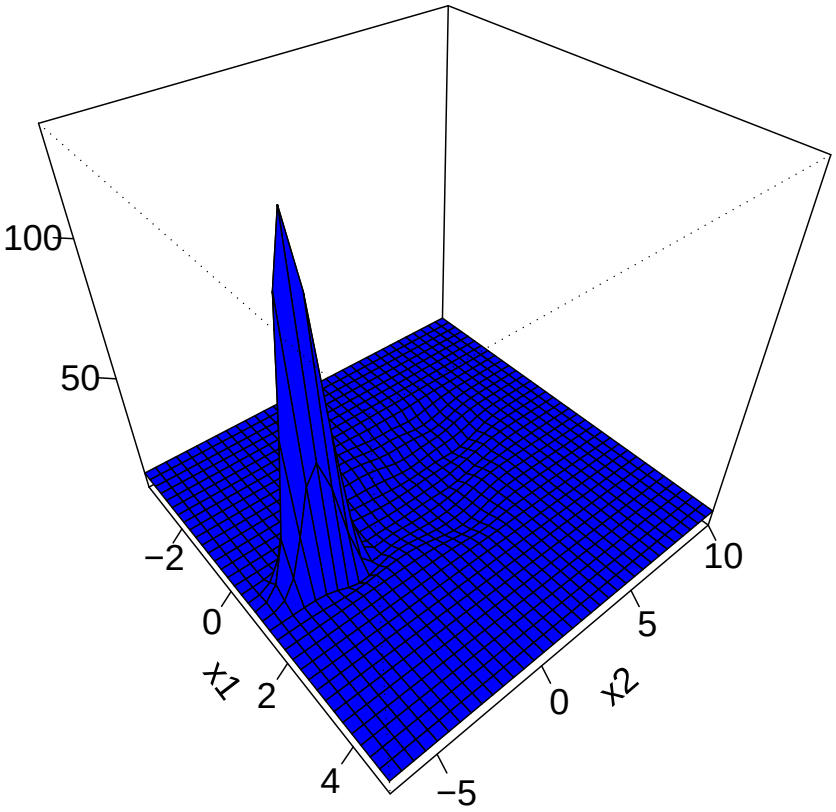
Sensitivity curve for $\hat{\theta}_2$.



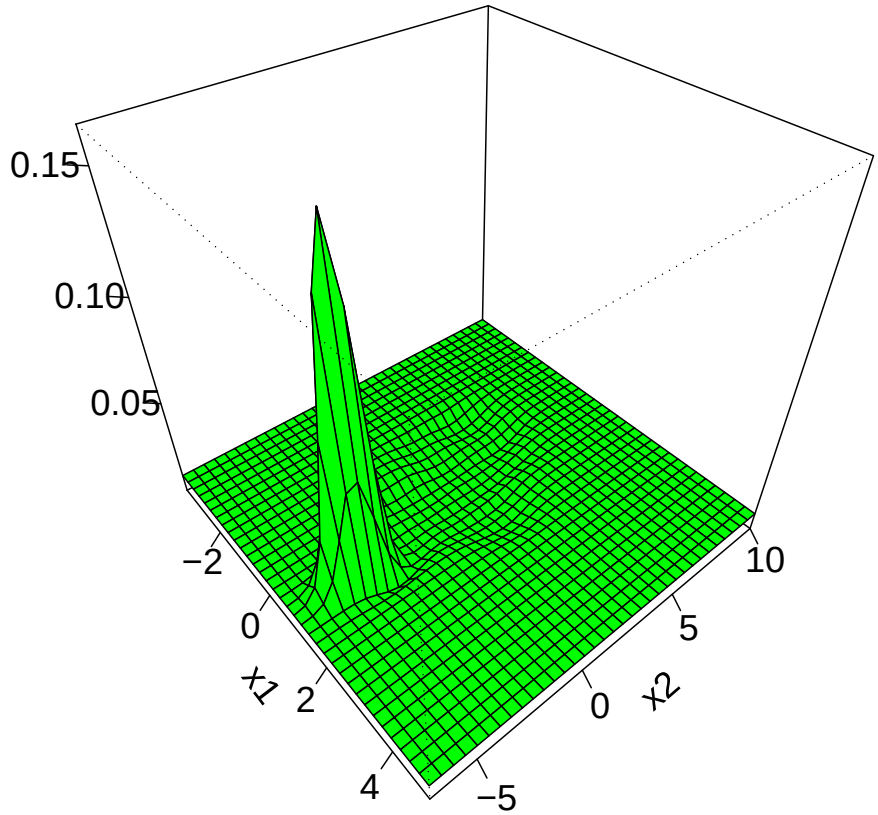
new obs at $z = (x_1, x_2, y = +1)$

KLR with RBF kernel

Sensitivity curve of $\hat{f}(x) + \hat{b}$



Difference of estimated event prob. $\Lambda(\hat{f}_{mod}(x) + \hat{b}_{mod}) - \Lambda(\hat{f}(x) + \hat{b})$



new obs at $x_1 = 2$, $x_2 = -2$, $y = +1$

4. SUMMARY

- kernel methods have many interesting properties for pattern recognition, estimation of class probabilities, regression, ...
- for pattern recognition:
  - uniform bounds for sensitivity curve for SVM, KLR, AdaBoost, ...
  - some robustness results for influence function
  - SVM with RBF kernel:
    $\longrightarrow$ smooth & local impact on $\hat{f}$
    $\longrightarrow$ bounded sensitivity curve
    $\longrightarrow$ smooth impact on $\hat{\theta}$, $\hat{b}$
    $\longrightarrow$ linear kernel seems to have global impact on $\hat{f}$

Current research: combination KLR + $\varepsilon-$SVR (or $\nu-$SVR)
joint work with: Dipl.-Stat. M. Marin-Galiano (DoMuS)

REFERENCES

Christmann, Steinwart (2003). *On robust properties of convex risk minimization methods for pattern recognition.* Submitted.

Christmann, Fischer, Joachims (2002). Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassifications. *Computational Statistics*, 17, 273-287.

Rousseeuw, Christmann (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis*, 43, 315-332.

Friedman, Hastie, Tibshirani (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Ann. Statist.*, 28, 337-407.

Hampel, Ronchetti, Rousseeuw, Stahel (1986). *Robust statistics: The Approach Based on Influence Functions*. Wiley.

Hastie, Tibshirani, Friedman (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer.

Höffgen, Simon, van Horn (1995). Robust Trainability of Single Neurons. *J. Computer and System Sciences*, 50, 114-125.

Schölkopf et al. (2000). New support vector algorithms. *Neural Computation*, 12, 1207-1245.

Schölkopf, Smola (2002) *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Smyth, Jørgensen (2002). Fitting Tweedie's Compound Poisson Model to Insurance Claims Data: Dispersion Modelling. *ASTIN Bulletin*, 32, 143-157.

Vapnik (1998). *Statistical Learning Theory*. Wiley.

Zhang (2001). Statistical behaviour and consistency of classification methods based on convex risk minimization. To appear in *Ann. Statist.*

christmann@statistik.uni-dortmund.de
www.statistik.uni-dortmund.de