

Exponential Families for Estimation

Alex J. Smola

Alex.Smola@anu.edu.au

NICTA, Statistical Machine Learning Program

Joint work with Yasemin Altun, Stephane Canu, Thomas Hofmann, Le Viet Quoc, and Vishy Vishwanathan

Outline

Estimation with Exponential Families

- Definition and examples
- Conditional models and missing variables
- General conditioning strategy

Applications

- Classification
- Novelty detection
- Regression
- Conditional Random Fields

Summary and Outlook

- Further applications
- Summary

The Exponential Family

Definition

A family of probability distributions which satisfy

$$p(x|\theta) = \exp(\langle \phi(x), \theta \rangle - g(\theta))$$

Cumulants Generating Function

$g(\theta)$ is the normalization for $p(x|\theta)$. It satisfies

$$\partial_{\theta} g(\theta) = \mathbf{E}_{x \sim p(x|\theta)} [\phi(x)] \quad \text{and} \quad \partial_{\theta}^2 g(\theta) = \mathbf{Cov}_{x \sim p(x|\theta)} [\phi(x)]$$

Details

- $\phi(x)$ is called the **sufficient statistic** of x .
- $g(\theta)$ is the **log-partition function** and it ensures that the distribution integrates out to 1. $g(\theta)$ is **convex**.
- Maximum Likelihood Estimation is a **convex** problem.
- Examples: Multinomial, Gaussian, Poisson, Laplace, Wishart, Dirichlet, Gamma, and Beta distribution

Example: Normal Distribution

Engineer's favorite

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \text{ where } x \in \mathbb{R} =: \mathcal{X}$$

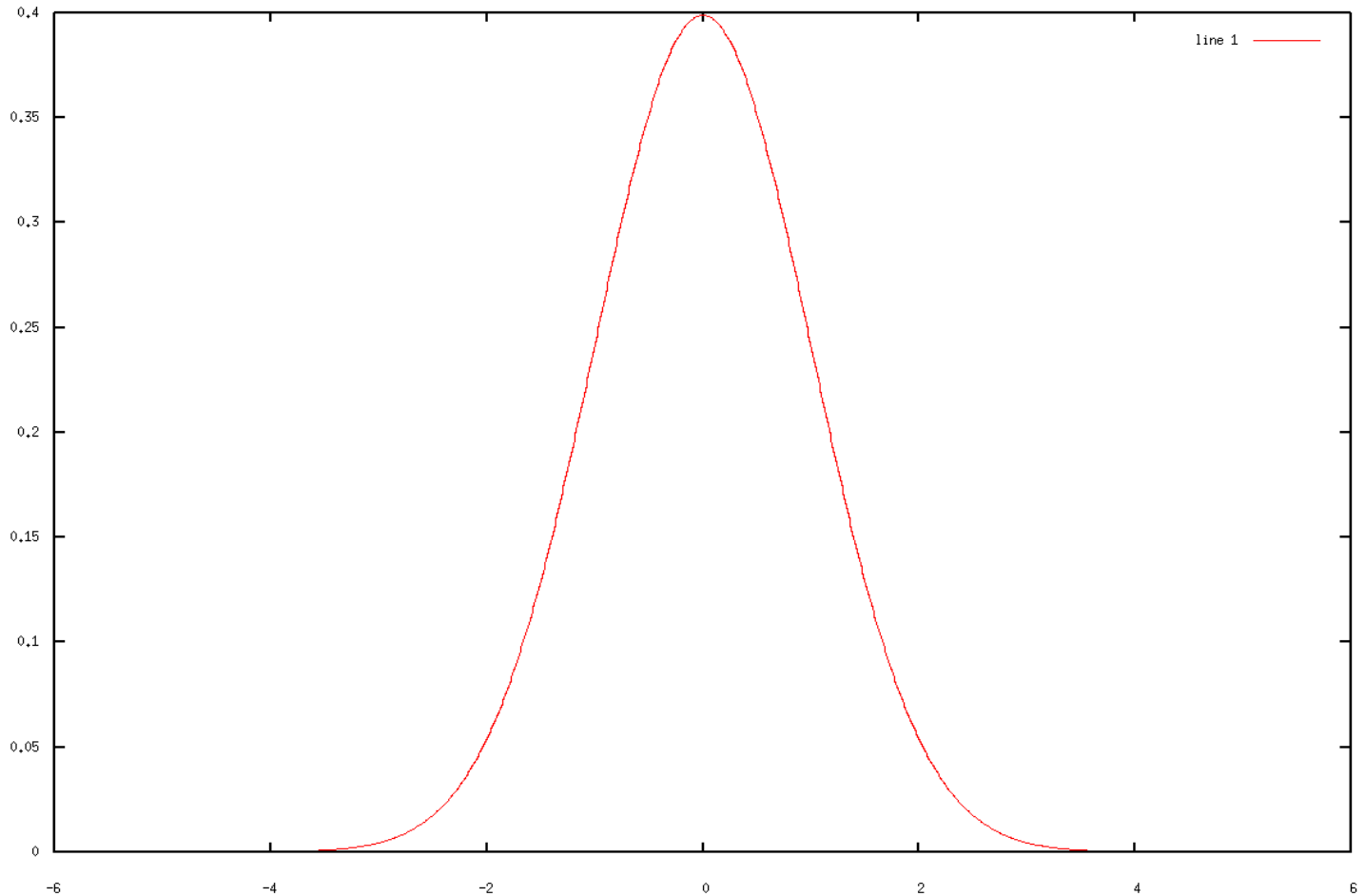
Massaging the math

$$p(x) = \exp\left(\underbrace{\langle (x, -0.5x^2), \theta \rangle}_{\phi(x)} - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2} \log(2\pi\sigma^2)\right)}_{g(\theta)}\right)$$

Using the substitution $\theta_2 := \sigma^{-2}$ and $\theta_1 := \mu\sigma^{-2}$ yields

$$g(\theta) = \frac{1}{2} [\theta_1^2 \theta_2^{-1} + \log 2\pi - \log \theta_2]$$

Example: Normal Distribution



Example: Multinomial Distribution

Many discrete events

Assume that we have n events, each which all may occur with a certain probability π_x .

Guessing the answer

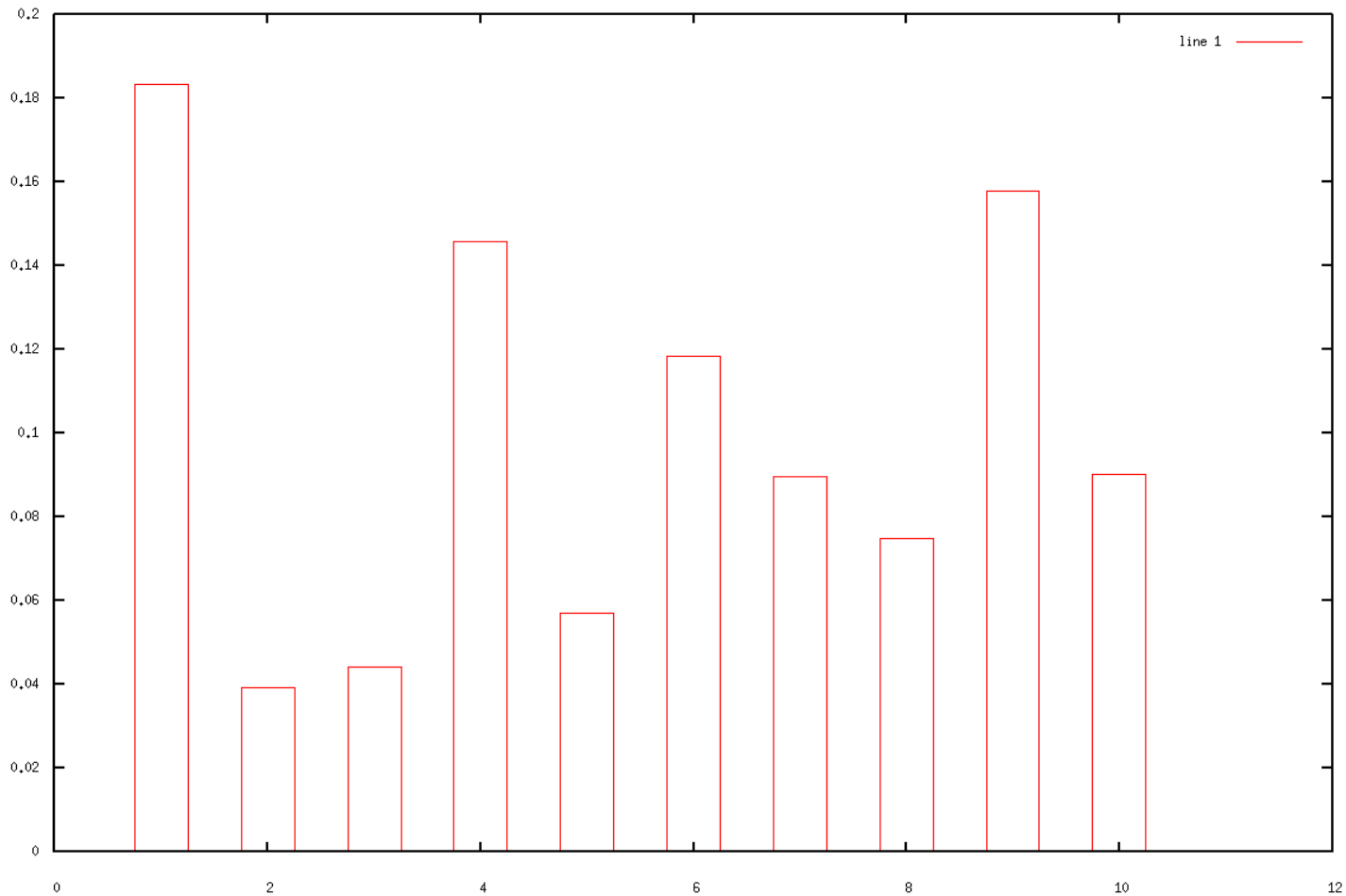
Use the map $\phi : x \rightarrow e_x$, that is, e_x is an element of the canonical basis $(0, \dots, 0, 1, 0, \dots)$ as sufficient statistic.

$$\implies p(x) = \exp(\langle e_x, \theta \rangle - g(\theta))$$

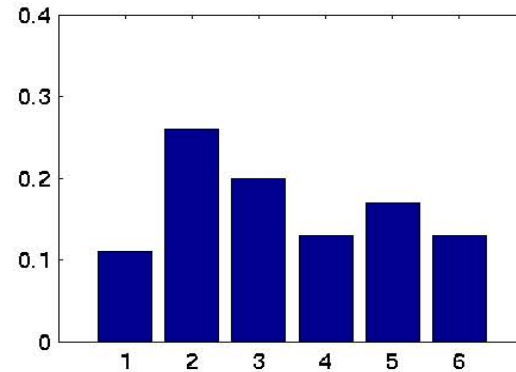
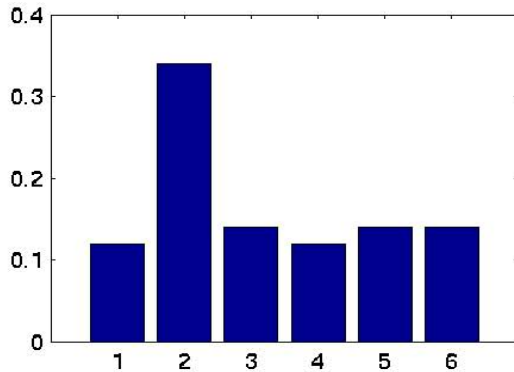
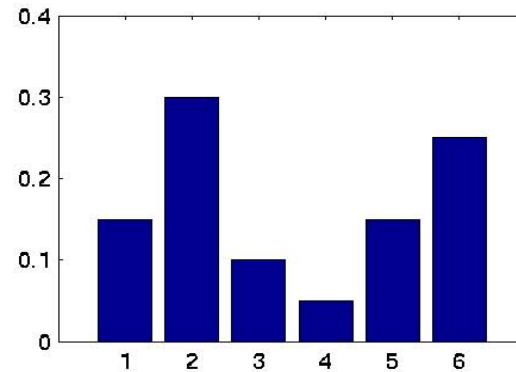
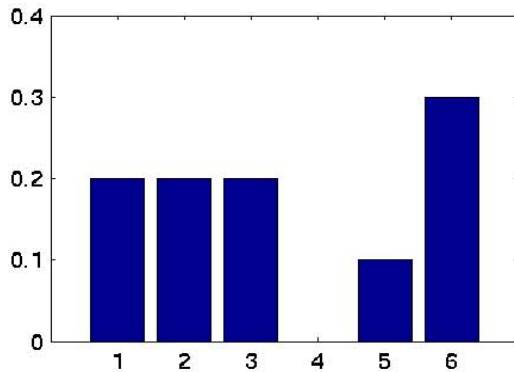
where the normalization is

$$g(\theta) = \log \sum_{i=1}^n \exp(\theta_i)$$

Example: Multinomial Distribution



Tossing a dice



Priors

Problems with Maximum Likelihood

With not enough data, parameter estimates will be bad.

Prior to the rescue

Often we know where the solution should be.

Normal Prior

Simply assume $\theta \sim \mathcal{N}(0, \sigma^2 \mathbf{1})$.

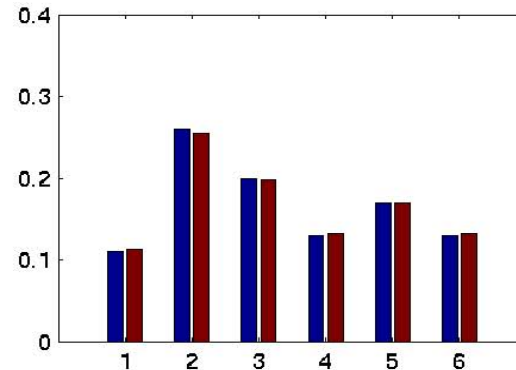
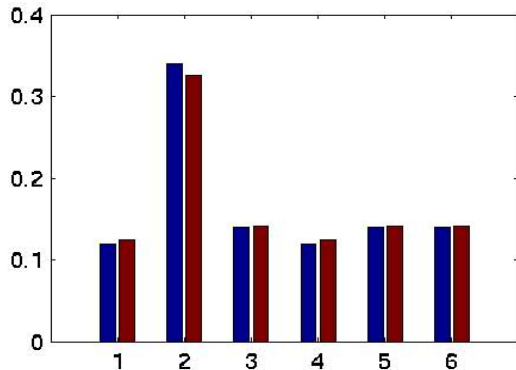
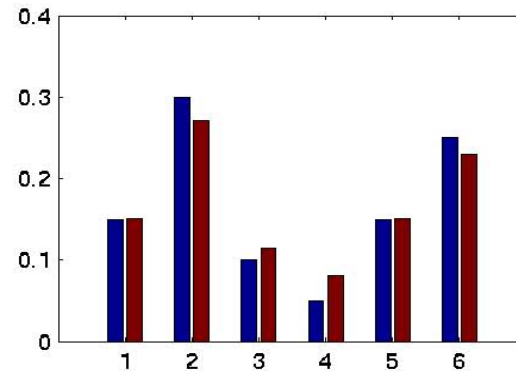
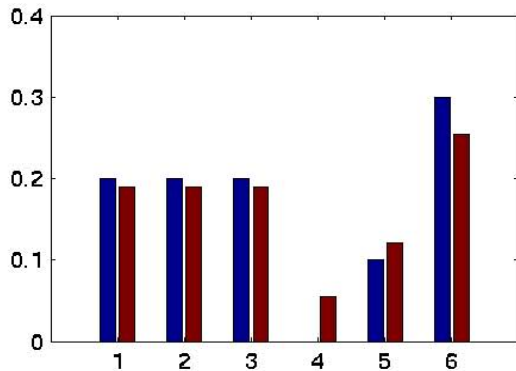
Posterior

$$-\log p(\theta|X) = \sum_{i=1}^m \underbrace{-\langle \phi(x_i), \theta \rangle + g(\theta)}_{-\log p(x_i|\theta)} + \underbrace{\frac{1}{2\sigma^2} \|\theta\|^2}_{-\log p(\theta)} + \text{const.}$$

Good News

Minimizing $-\log p(\theta|X)$ is a **convex** optimization problem.

Tossing a dice with priors



Conditional Distributions

Conditional Density and MAP Estimation

$$p(y|x, \theta) = \exp(\langle \phi(x, y), \theta \rangle - g(\theta|x))$$
$$-\log p(\theta|X, Y) = \sum_{i=1}^m -\langle \phi(x_i, y_i), \theta \rangle + g(\theta|x_i) + \frac{1}{2\sigma^2} \|\theta\|^2 + c$$

Solving the Problem

Expand θ in a linear combination of $\phi(x_i, y_i)$ and solve convex problem in expansion coefficients.

Missing (Latent) Variables

$x = (x^u, x^o)$ where x^o observed and x^u unobserved.

$$p(y|x^o, \theta) = \int p(y, x^u|x^o, \theta) dx^u = \exp(g(\theta|x^o, y) - g(\theta|x^o))$$

The likelihood $p(y|x^o, \theta)$ is no longer log-concave!

General Strategy

Choose a suitable sufficient statistic $\phi(x, y)$

- Conditionally multinomial distribution leads to Gaussian Process multiclass estimator: we have a distribution over n classes which depends on x .
- Conditionally Gaussian leads to Gaussian Process regression: we have a normal distribution over a random variable which depends on the location.
- Structured $\phi(x, y)$ leads to Conditional Random Fields.

Kernel Trick MK II

Rewrite optimization problem in terms of

$$k((x, y), (x', y')) = \langle \phi(x, y), \phi(x', y') \rangle$$

Gaussian Process Classification

Binomial Model

$$\phi(x, y) = y\phi(x) \text{ where } y \in \{\pm 1\}$$

leads to standard GP classification problem.

Optimization Problem

Minimize the negative log-posterior, that is, solve

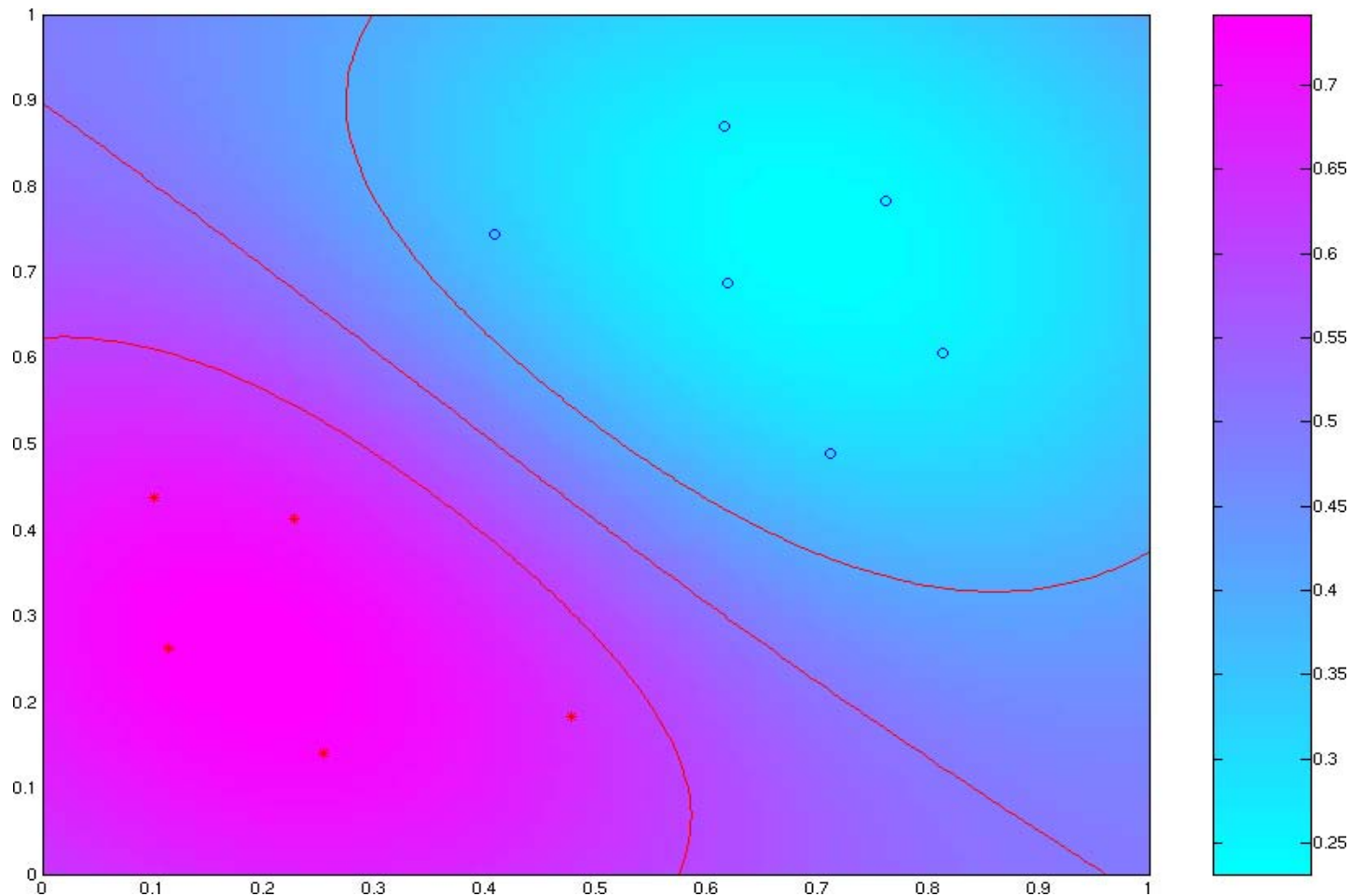
$$\underset{\theta}{\text{minimize}} \sum_{i=1}^m \underbrace{-\log p(y_i|x_i, \theta)}_{g(\theta|x_i) - \langle \phi(x_i, y_i), \theta \rangle} \underbrace{-\log p(\theta)}_{\frac{1}{2\sigma^2}\|\theta\|^2}$$
$$g(\theta|x_i^o) - g(\theta|x_i^o, y_i)$$

Support Vector Classification

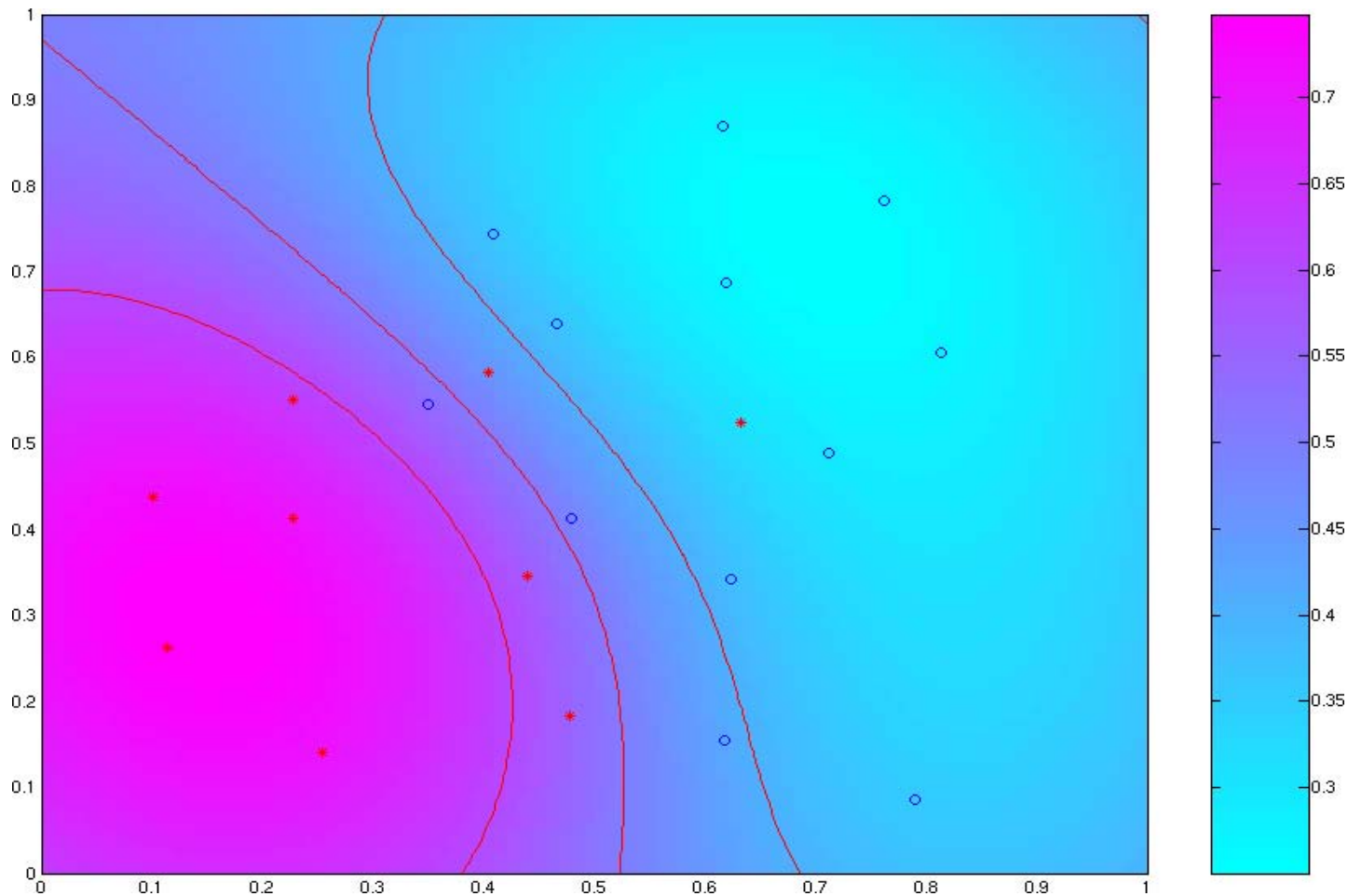
Compute margin as log-likelihood ratio via

$$\rho(x, y, \theta) := \log \frac{p(y|x, \theta)}{\max_{\tilde{y} \neq y} p(\tilde{y}|x, \theta)}$$

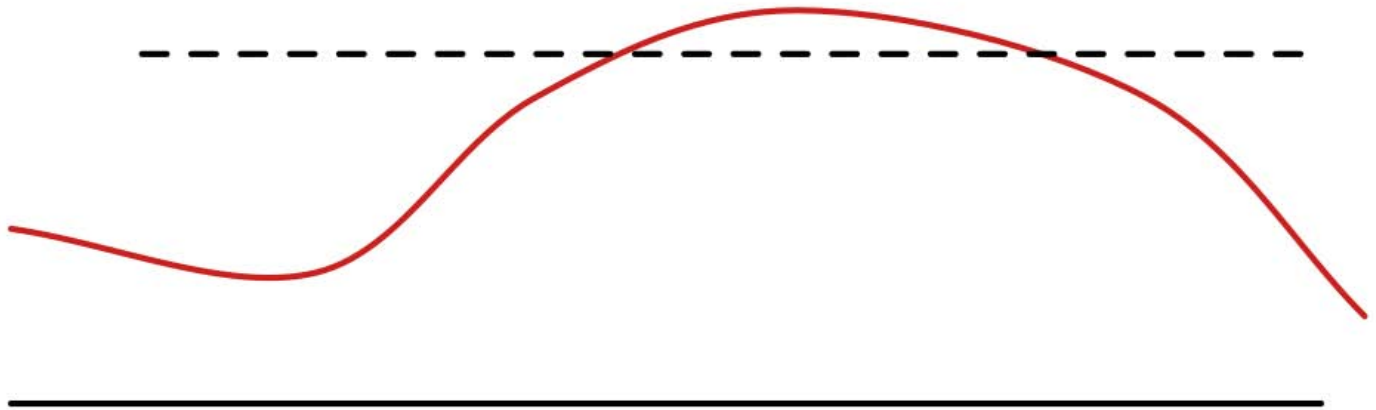
A Toy Example



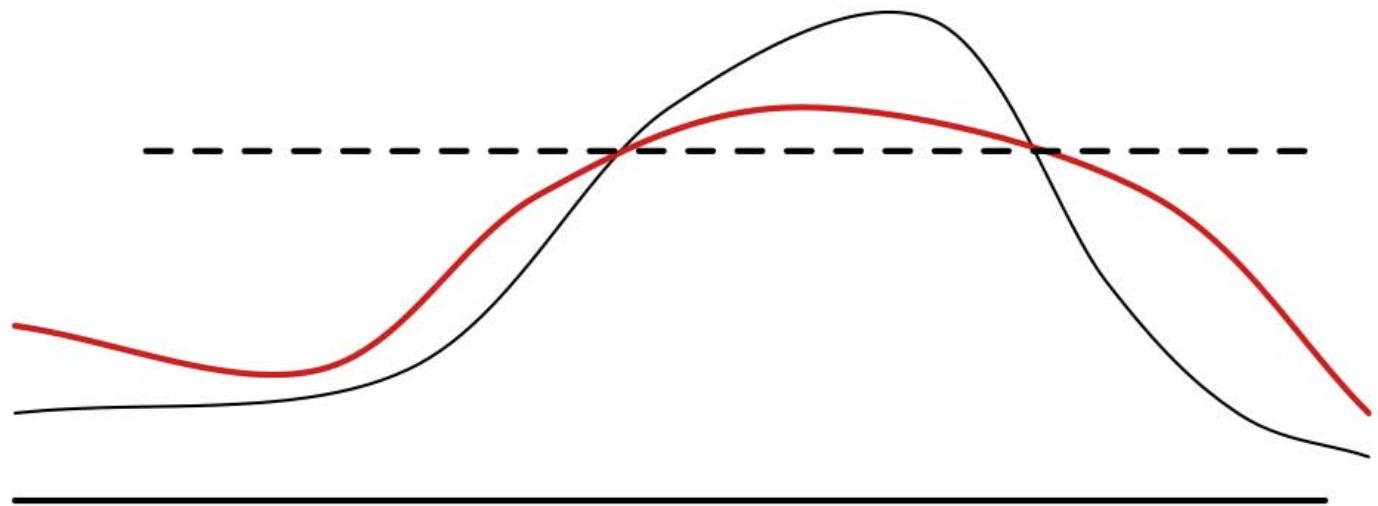
Noisy Data



Novelty Detection



Novelty Detection



Novelty Detection

Conventional Approach

- Estimate density, e.g. via penalized log-likelihood
- Declare all x with $p(x|\theta) < \epsilon$ as novel

Problems

- Estimator spends capacity on high density regions
- We don't need normalized density

Solution: Trimmed Log-likelihood

Use unnormalized density $\frac{p(x|\theta)}{\exp(g(\theta))}$, threshold at $\exp(\rho)$

Optimization Problem

Minimize trimmed log-likelihood, penalized by $-\log p(\theta)$

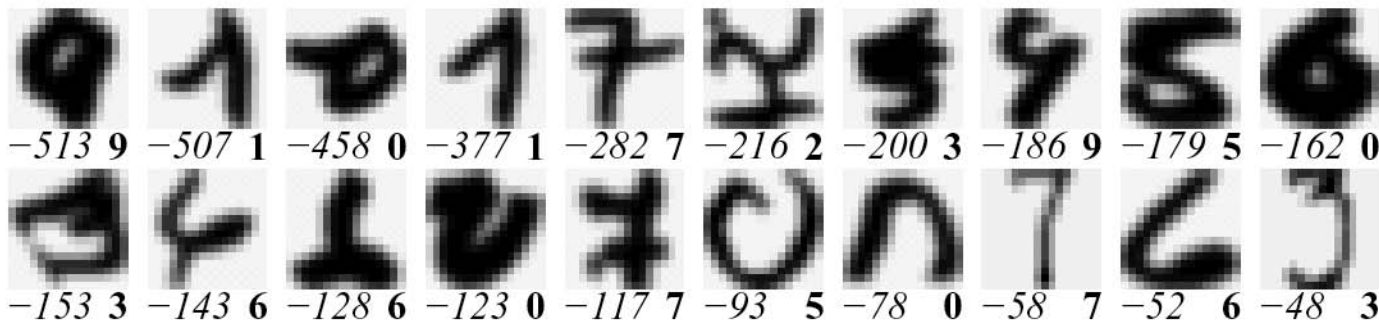
$$\underset{\theta}{\text{minimize}} \sum_{i=1}^m \max(0, \rho + \underbrace{g(\theta) - \log p(x_i|\theta)}_{\substack{-\langle \phi(x_i), \theta \rangle \\ -g(\theta|x_i^o)}}) + \frac{1}{2\sigma^2} \|\theta\|^2$$

USPS Digits

Random Digits

3 4 8 6 1 1 3 6
0 0 4 7 1 4 4 2
6 0 4 3 3 7 4 1
3 5 0 0 2 1 0 0
1 7 9 2 0 6 0 0

Worst Digits



Heteroscedastic GP Regression

Sufficient Statistic

We pick $\phi(x, y) = (y\phi_1(x), y^2\phi_2(x))$, that is

$k((x, y), (x', y')) = k_1(x, x')yy' + k_2(x, x')y^2y'^2$ where $y, y' \in \mathbb{R}$

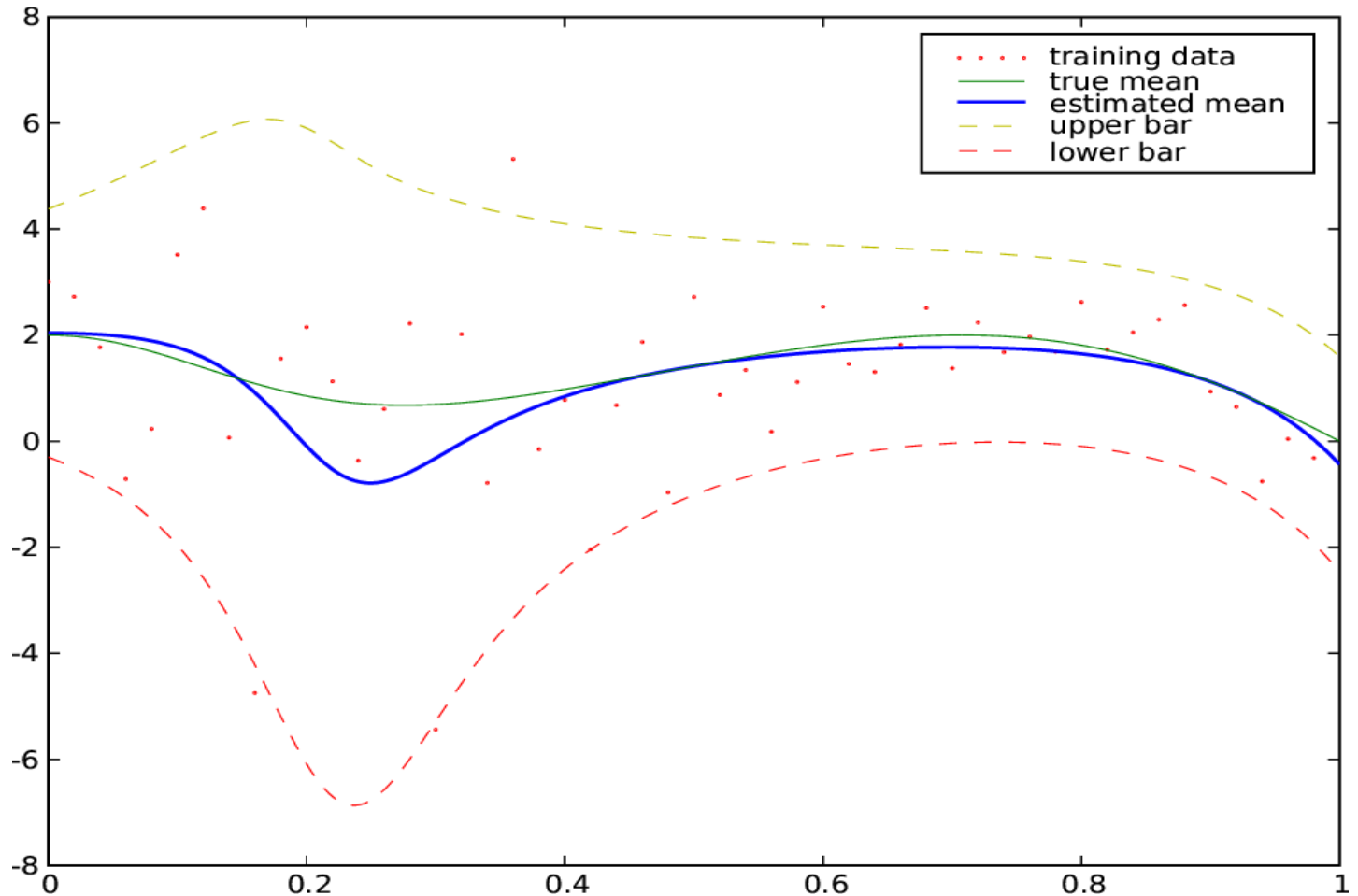
Hence estimate mean and variance **simultaneously**.

Optimization Problem

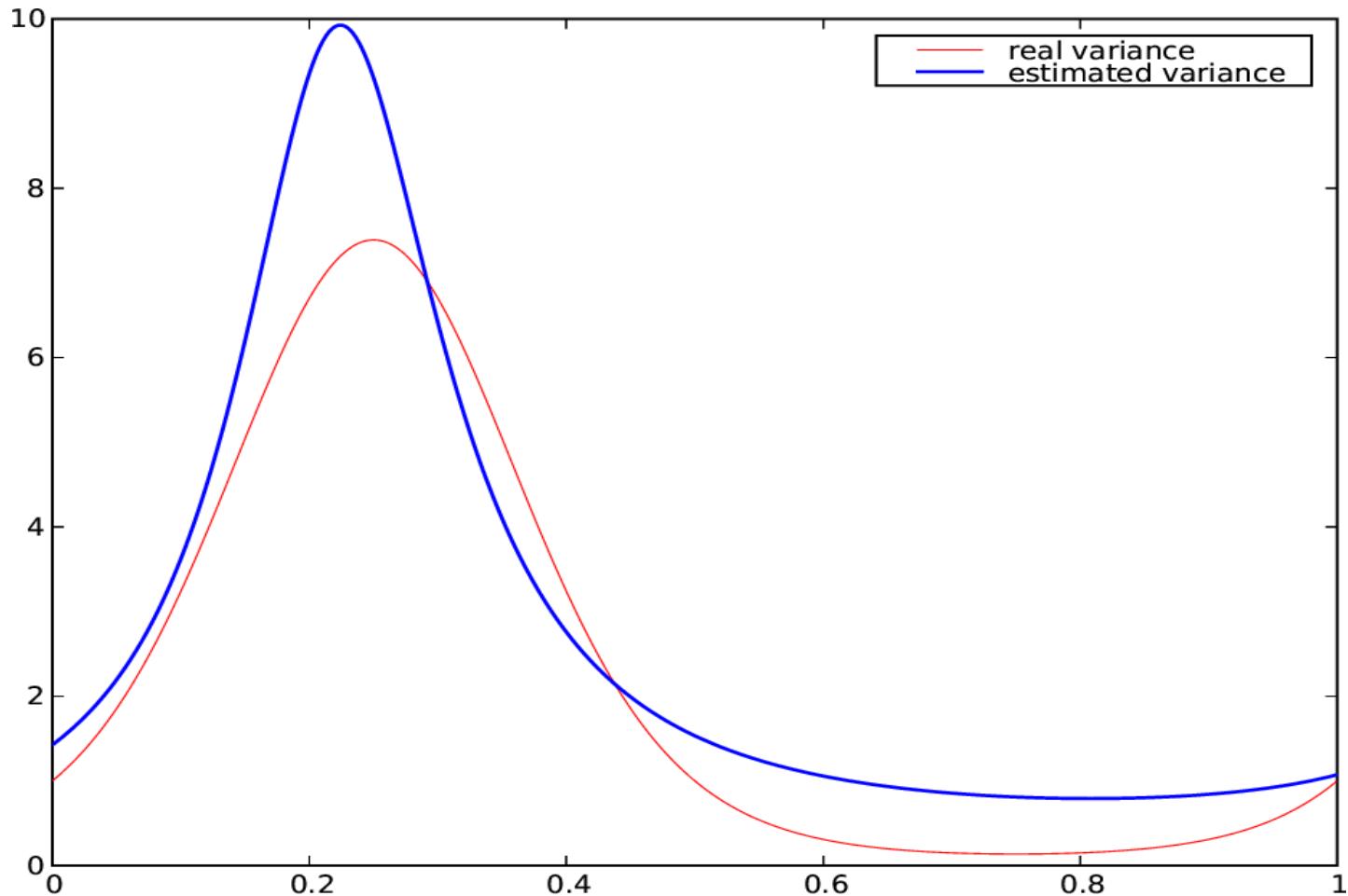
$$\begin{aligned} \text{minimize } & \sum_{i=1}^m \left[-\frac{1}{4} \left[\sum_{j=1}^m \alpha_{1j} k_1(x_i, x_j) \right]^\top \left[\sum_{j=1}^m \alpha_{2j} k_2(x_i, x_j) \right]^{-1} \left[\sum_{j=1}^m \alpha_{1j} k_1(x_i, x_j) \right] \right. \\ & \left. -\frac{1}{2} \log \det -2 \left[\sum_{j=1}^m \alpha_{2j} k_2(x_i, x_j) \right] - \sum_{j=1}^m \left[y_i^\top \alpha_{1j} k_1(x_i, x_j) + (y_i^\top \alpha_{2j} y_j) k_2(x_i, x_j) \right] \right] \\ & + \frac{1}{2\sigma^2} \sum_{i,j} \alpha_{1i}^\top \alpha_{1j} k_1(x_i, x_j) + \text{tr} \left[\alpha_{2i} \alpha_{2j}^\top \right] k_2(x_i, x_j). \\ \text{subject to } & 0 \succ \sum_{i=1}^m \alpha_{2i} k(x_i, x_j) \end{aligned}$$

- The problem is convex
- The log-determinant from the normalization of the Gaussian acts as a **barrier function**, i.e. a nice SDP.

Heteroscedastic Regression



Variance Estimate



Computational Issues

Newton Method with CG Solver

Use Newton method to compute update direction, CG solver instead of inverting Hessian.

Lazy Evaluation

Never build explicit Hessian.

Reduced Rank

Use incomplete Cholesky factorization for low-rank approximation.

Result

m	100	200	500	1k	2k	5k	10k	20k
Direct Hessian	8	18	90	607	3551	-	-	-
Hessian vector	9	15	38	115	752	-	-	-
Reduced rank	7	7	12	30	54	179	368	727

This yields scaling of $O(m^{2.1})$, $O(m^{1.4})$, and $O(m^{0.95})$.

Spatial Interpolation Comparison'04

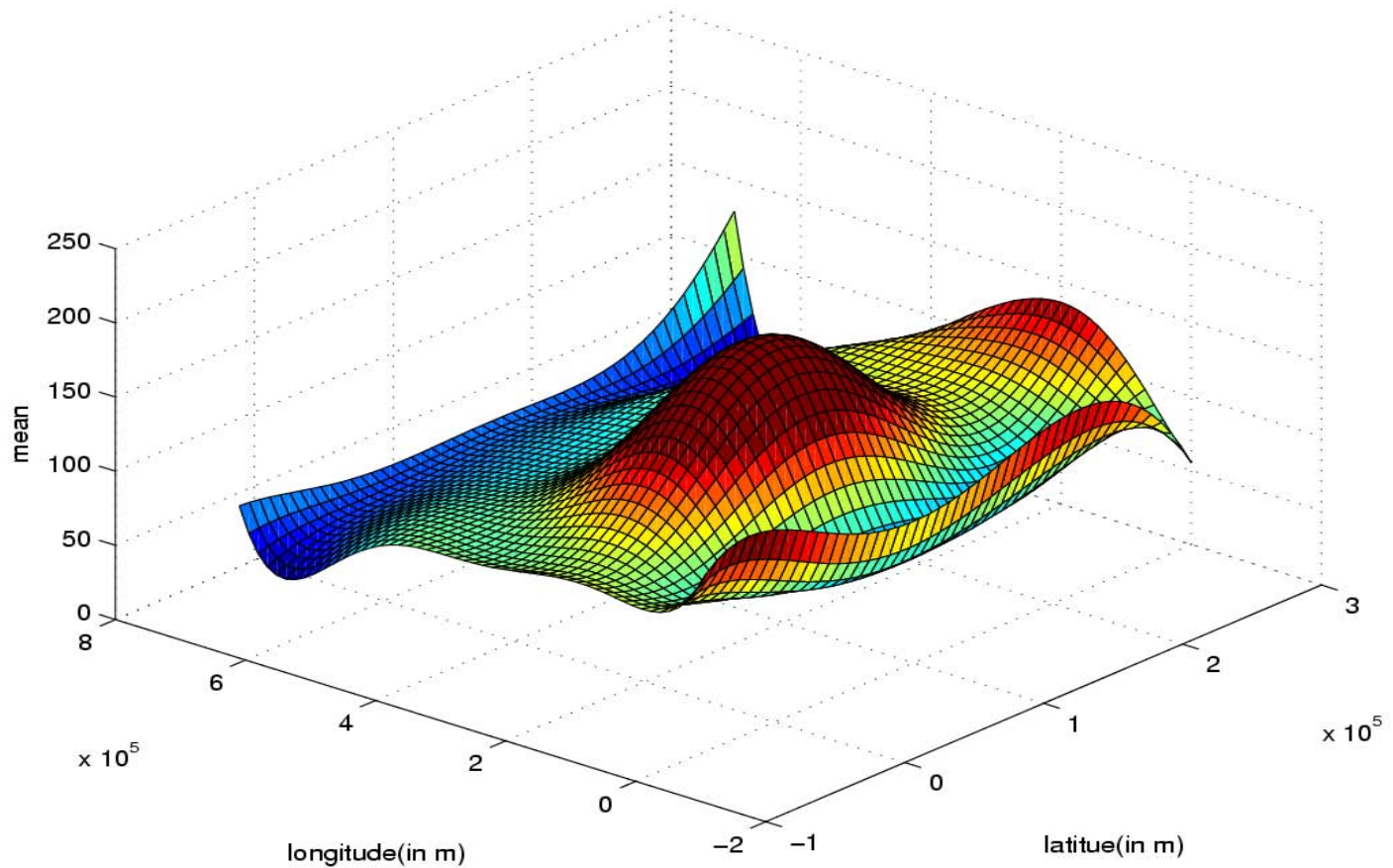
Dataset

- Ten days of normal radiation measurements (prior knowledge)
- Two days for which prediction of radiation is required. Scenario 1 is normal, scenario 2 models nuclear disaster (localized high radiation event).

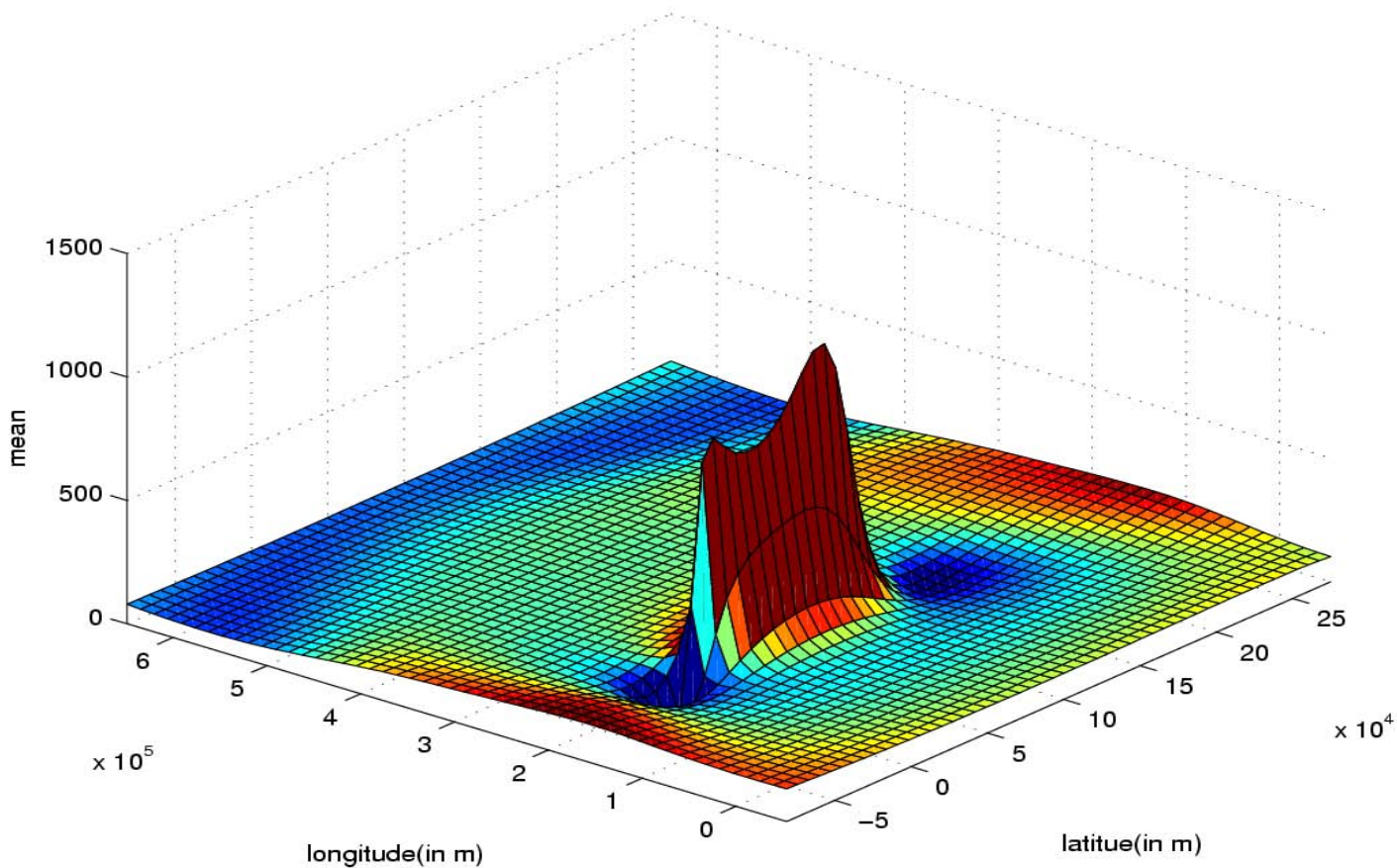
Performance

	GP Regression	HGP Regression
Normal day	13.3 ± 0.4	12.8 ± 0.6
Anomaly	86.8 ± 6.5	49.4 ± 6.5

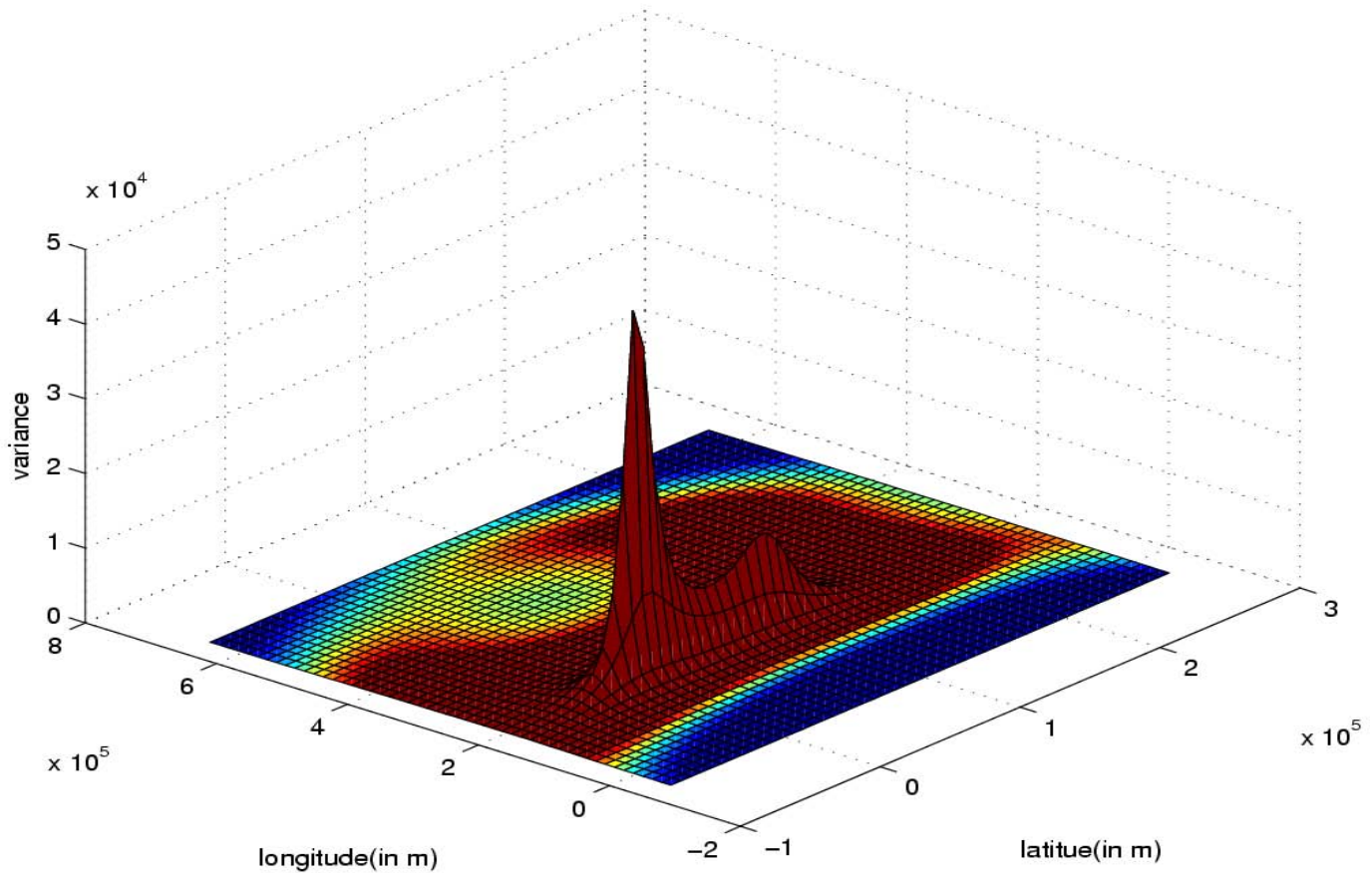
Standard Gaussian Process



Heteroscedastic GP (mean)

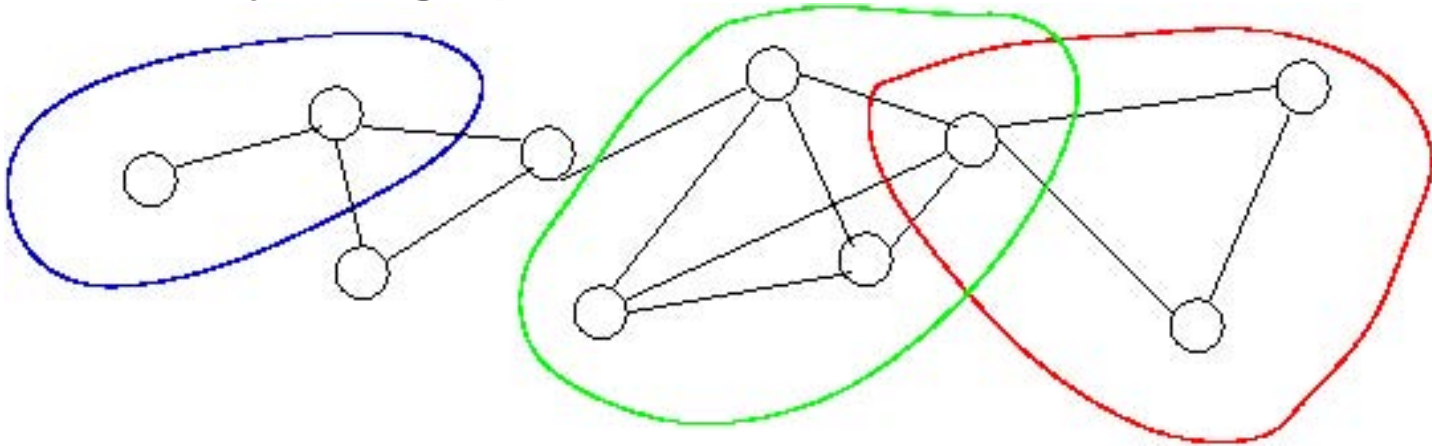


Heteroscedastic GP (variance)



Structured Observations

Joint density and graphical models



Hammersley-Clifford Theorem

$$p(x) = \frac{1}{Z} \exp \left(\sum_{c \in \mathcal{C}} \psi_c(x_c) \right)$$

Decomposition of any $p(x)$ into product of potential functions on maximal cliques.

Application to Exponential Families

Hammersley-Clifford Corollary

Combining the CH-Theorem and exponential families

$$p(x) = \frac{1}{Z} \exp \left(\sum_{c \in \mathcal{C}} \psi_c(x_c) \right)$$

$$p(x) = \exp (\langle \phi(x), \theta \rangle - g(\theta))$$

we obtain a decomposition of $\phi(x)$ into

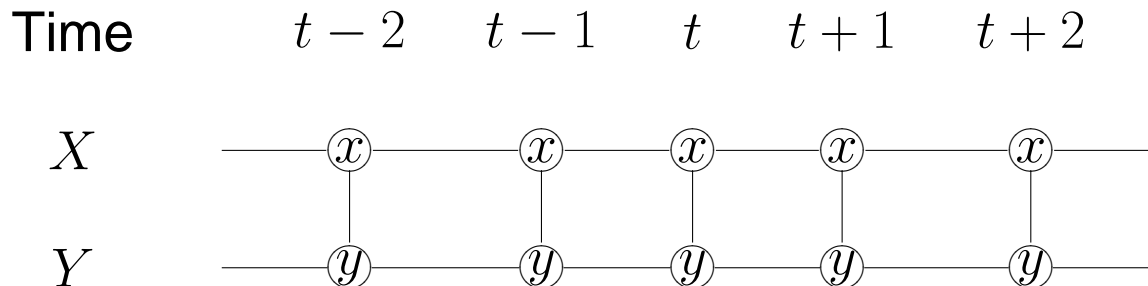
$$p(x) = \exp \left(\sum_{c \in \mathcal{C}} \langle \phi_c(x_c), \theta_c \rangle - g(\theta) \right)$$

Consequence for Kernels

$$k(x, x') = \sum_{c \in \mathcal{C}} k_c(x_c, x'_c)$$

Conditional Random Fields

Dependence structure between variables



Key Points

- We can drop cliques in x : they do not affect $p(y|x, \theta)$.
- Compute $g(\theta|x)$ via dynamic programming.
- Assume stationarity of the model, that is θ_c does not depend on the position of the clique.
- We only need a sufficient statistic $\phi_{xy}(x_t, y_t)$ and $\phi_{yy}(y_t, y_{t+1})$.

Computational Issues

Conditional Probabilities:

$$p(y|x, \theta) \propto \prod_{t=1}^T \underbrace{\exp(\langle \phi_{xy}(x_t, y_t), \theta_{xy} \rangle + \langle \phi_{yy}(y_t, y_{t+1}), \theta_{yy} \rangle)}_{M(y_t, y_{t+1})}$$

So we can compute $p(y_t|x, \theta)$ and $p(y_t, y_{t+1}|x, \theta)$ via dynamic programming.

Objective Function:

$$-\log p(\theta|X, Y) = \sum_{i=1}^m -\langle \phi(x_i, y_i), \theta \rangle + g(\theta|x_i) + \frac{1}{2\sigma^2} \|\theta\|^2 + c$$

$$\partial_{\theta} -\log p(\theta|X, Y) = \sum_{i=1}^m -\phi(x_i, y_i) + \mathbf{E}[\phi(x_i, y_i)|x_i] + \frac{1}{\sigma^2} \theta$$

We only need $\mathbf{E}[\phi_{xy}(x_{it}, y_{it})|x_i]$ and $\mathbf{E}[\phi_{yy}(y_{it}, y_{i(t+1)})|x_i]$.

Some Applications

Named Entity Tagging

Annotate words in documents based on their function (e.g. name, verb, location, ...). This relies on neighborhood information.

Pitch Accent Prediction

Estimate accent from dialog (useful for audio dialog systems). Again needs local correlations between words.

Functional Annotation in Gene Sequences

Relative position of Intron and Exon on sequence matters. Can use grammar in annotation process.

Parts Based Models of Recognition

Faces are composed of nose, mouth, eyes, ears, etc. So we can build detectors on each of them and combine them in a graphical model. Inference is nontrivial.

Further Extensions and Applications

Spatial Poisson Models

Use models with conditionally Poisson distribution.

Multiclass Perceptron

Use multiclass margin instead of binary margin.

Bayesian Semi-supervised Learning

We only observe some of the labels.

Clustering

We observe no labels — MAP label assignment.

Maximum Margin Markov Nets

Margin definition for structured nets.

Texture Synthesis

Controlled nonparametric interaction models

Image Denoising and Superresolution

Learn interactions and fill in.

Shameless Plugs

We are hiring. For details contact

- Alex.Smola@nicta.com.au (<http://www.nicta.com.au>)

Positions

- PhD scholarships
- Postdoctoral positions, Senior researchers
- Long-term visitors (sabbaticals etc.)

More details on kernels

- <http://www.kernel-machines.org>
- <http://www.learning-with-kernels.org>
Schölkopf and Smola: Learning with Kernels

Machine Learning Summer Schools

MLSS'05 Chicago, USA, June 2005