Non-parametric Bayesian Methods in Machine Learning

Dr. Simon Rogers
School of Computing Science
University of Glasgow
simon.rogers@glasgow.ac.uk
@sdrogers

May 10, 2014

Outline

- (My) Bayesian philosophy
- Gaussian Processes for Regression and Classification (Monday)
    - GP preliminaries
    - *Application 1*: typing on touch-screens
    - Classification (including semi-supervised)
    - *Application 2*: clinical (dis)-agreement
- Dirichlet Process flavoured Cluster Models (Tuesday)
    - DP preliminaries
    - *Application 3*:Idenfitying metabolites
    - *Application 4*:Cluster models for multiple data views
- Summary

Relevant publications

- The four applications are described in the following papers:
  - Uncertain Text Entry on Mobile Devices Weir et. al, CHI 2014
  - Investigating the Disagreement Between Clinicians' Ratings of Patients in ICUs Rogers et. al 2013, IEEE Trans Biomed Health Inform
  - MetAssign: Probabilistic annotation of metabolites from LC–MS data using a Bayesian clustering approach Daly et. al, Bioinformatics, under review
  - Infinite factorization of multiple non-parametric views Rogers et. al, Machine Learning 2009

About me

- I'm not a statistican by training (don't ask me to prove anything!).
- Education:
  - Undergraduate Degree: Electrical and Electronic Engineering (Bristol)
  - PhD: Machine Learning Techniques for Microarray Analysis (Bristol)
- Currently:
  - Lecturer: Computing Science
  - Research Interests: Machine Learning and Applied Statistics in Computational Biology and Human-Computer Interaction (HCI)

Lecture 4: GPs for classification and ordinal regression via the auxiliary variable trick

Dr. Simon Rogers
School of Computing Science
University of Glasgow
simon.rogers@glasgow.ac.uk
@sdrogers

May 10, 2014

GPs for Classification and ordinal regression

- What if our observation model is non-Gaussian?
  - Classification:

$$P(y_n = 1 | f_n) = \int_{-\infty}^{f_n} \mathcal{N}(z|0,1)\ dz = \phi(f_n)$$

  - Logistic Regression:

$$P(y_n = k | f_n) = \phi(b_{k+1}) - \phi(b_k)$$

  - etc
- Analytical inference is no longer possible
- I'll cover how to do inference in these models and extensions with the *auxiliary variable trick*

Binary classification

- Problem setup: we observe N data / target pairs $(\mathbf{x}_n, y_n)$ where $y_n \in \{0, 1\}$
- Place a GP prior on a set of latent variables f_n

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$$

- Use the probit likelihood:

$$P(y_n = 1 | f_n) = \phi(f_n) = \int_{-\infty}^{f_n} \mathcal{N}(z | 0, 1)\ dz$$

- Inference in this form is hard

Auxiliary Variable Trick

- Re-write the probit function:

$$
\begin{aligned}
P(y_n = 1 | f_n) &= \int_{-\infty}^{f_n} N(z|0,1)\ dz \\
&= \int_{-\infty}^{0} N(z|-f_n,1)\ dz \\
&= \int_{0}^{\infty} N(z|f_n,1)\ dz \\
&= \int_{-\infty}^{\infty} \delta(z>0)\mathcal{N}(z|f_n,1)\ dz
\end{aligned}
$$

where $\delta(expr)$ is 1 if *expr* is true, and 0 otherwise.

Auxiliary Variable Trick

- If we define $P(y_n = 1|z_n) = \delta(z_n > 0)$ then we have:

$$P(y_n = 1|f_n) = \int_{-\infty}^{\infty} P(y_n = 1|z_n) p(z_n|f_n)\ dz_n$$

- and could therefore remove the integral to obtain a model including z_n:

$$p(y_n = 1, z_n|f_n) = P(y_n = 1|z_n) p(z_n|f_n)$$

- Doing inference in this model (i.e. with additional variables z_n) is much easier (but still not analytically tractable)
- Note: $P(y_n = 0|z_n) = \delta(z_n < 0)$

Example - Gibbs sampling for binary classification

- An easy way to perform inference in the augmented model is via Gibbs sampling
- Sample $z_n | f_n, y_n$:

$$
\begin{aligned}
p(z_n | f_n, y_n = 0) &\propto \delta(z_n < 0)\mathcal{N}(z_n | f_n, 1) \\
p(z_n | f_n, y_n = 1) &\propto \delta(z_n < 1)\mathcal{N}(z_n | f_n, 1)
\end{aligned}
$$

Example - Gibbs sampling for binary classification

- An easy way to perform inference in the augmented model is via Gibbs sampling
- Sample $z_n | f_n, y_n$:

$$
\begin{aligned}
p(z_n | f_n, y_n = 0) &\propto \delta(z_n < 0)\mathcal{N}(z_n | f_n, 1) \\
p(z_n | f_n, y_n = 1) &\propto \delta(z_n < 1)\mathcal{N}(z_n | f_n, 1)
\end{aligned}
$$

- Sample $\mathbf{f} | \mathbf{z}, \mathbf{C}$

$$p(\mathbf{f} | \mathbf{z}, \mathbf{C}) = \mathcal{N}(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$$

where

$$\boldsymbol{\Sigma}_f = \left(\mathbf{I} + \mathbf{C}^{-1}\right)^{-1}, \quad \boldsymbol{\mu}_f = \boldsymbol{\Sigma}_f^{-1}\mathbf{z}$$

- Repeat ad infinitum

Example - Gibbs sampling for binary classification

- To make predictions:
  - At each sampling step, do a (noise-free) GP regression using the current sample of $\mathbf{f}$ to get a density over f_* (Details in a previous slide).
  - Sample a specific realisation of f_* from this density.
  - Compute $\phi(f_*)$ (or sample a z_* and then record whether it's > 0 or not)
  - Average this value over all Gibbs sampling iterations!

Example - binary classification

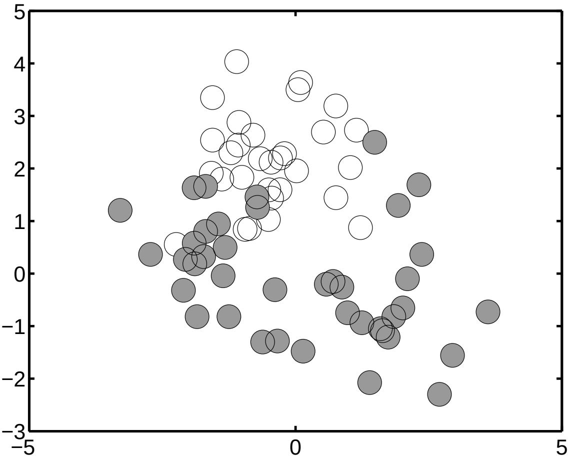


Figure 14 : Some simple classification data

Example - binary classification

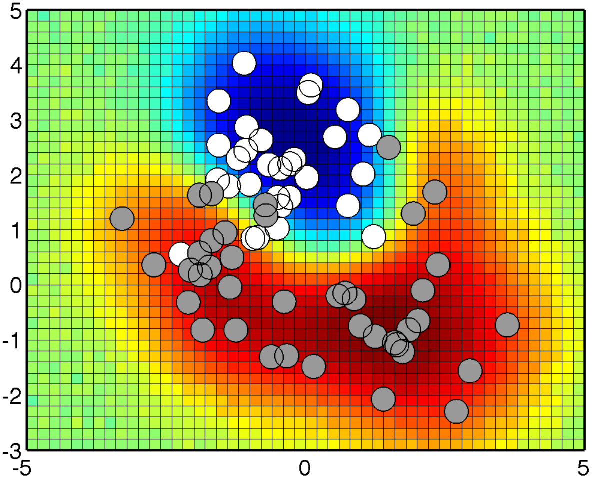


Figure 15 : Predictive probabilities averaged over 1000 Gibbs samples using an RBF covariance. As γ is increased, the model overfits.

Example - binary classification

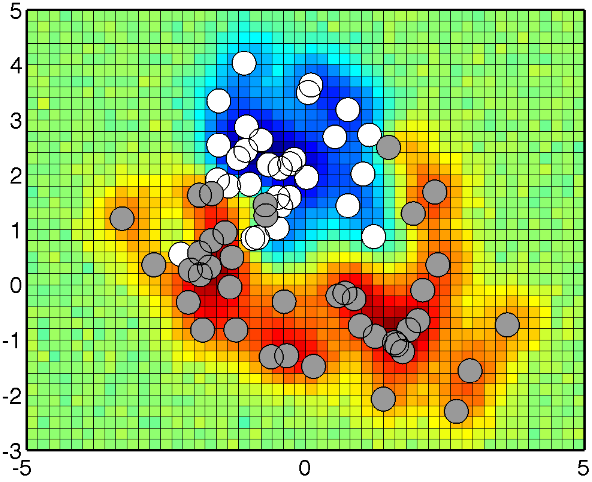


Figure 15 : Predictive probabilities averaged over 1000 Gibbs samples using an RBF covariance. As γ is increased, the model overfits.

Example - binary classification

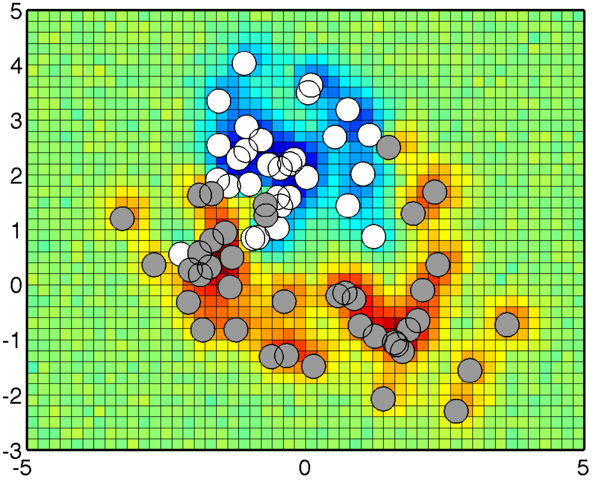


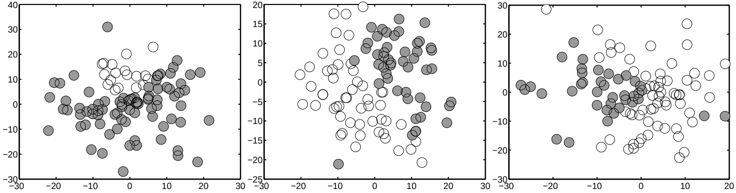
Figure 15 : Predictive probabilities averaged over 1000 Gibbs samples using an RBF covariance. As γ is increased, the model overfits.

Note

- Inference:
  - Gibbs sampling isn't the only option
  - A popular alternative is Variational Bayes

Note 2 – The Generative Process

- Sometimes it's useful to think of the generative process defined by the model.
- In this case, to generate N values of y_n given the associated x_n:
  - Sample $\mathbf{f}$ from a GP with mean $\mathbf{0}$ and Covariance matrix $\mathbf{C}$.
  - For each $n = 1 \ldots N$:
    - Sample $z_n \sim \mathcal{N}(f_n, 1)$
    - If $z_n > 0$ set $y_n = 1$, otherwise $y_n = 0$.
- Some examples:



GP classification exercise

TASK [2]

- Explore GP binary classification with auxiliary variables using `gp_class_task.m`
- Try:
  - Generating data from different distributions
  - Varying covariance function and parameters
  - Taking more posterior samples
- You will also need `plotClassdata.m` and `kernel.m`

A more general idea

- Models of this form:
  - $\mathbf{f} \sim GP$
  - $z_n \sim \mathcal{N}(f_n, 1)$
  - $P(y_n|z_n) = \delta(f(z_n))$
- Can be used for more than just binary classification.

A more general idea

- Models of this form:
  - $\mathbf{f} \sim GP$
  - $z_n \sim \mathcal{N}(f_n, 1)$
  - $P(y_n|z_n) = \delta(f(z_n))$
- Can be used for more than just binary classification.
- Ordinal Regression:
  - $P(y_n = k|z_n)$ is now chopped at both ends:

$$P(y_n = k|z_n) = \delta(b_k < z_n < b_{k+1})$$

  - Gibbs distribution for z_n therefore involves a Gaussian truncated at both ends.

A more general idea

- Models of this form:
  - $\mathbf{f} \sim GP$
  - $z_n \sim \mathcal{N}(f_n, 1)$
  - $P(y_n|z_n) = \delta(f(z_n))$
- Can be used for more than just binary classification.
- Ordinal Regression:
  - $P(y_n = k|z_n)$ is now chopped at both ends:

$$P(y_n = k|z_n) = \delta(b_k < z_n < b_{k+1})$$

  - Gibbs distribution for z_n therefore involves a Gaussian truncated at both ends.
- As well as multi-class and semi-supervised classification...

Multi-class classification

- The previous treatment can be extended to multiple classes.
- For a problem with K classes:
  - K GP priors, K N-dimensional latent vectors $\mathbf{f}_k$.
  - $N \times K$ auxiliary variables $z_{nk} \sim \mathcal{N}(f_{nk}, 1)$
  - And:

$$P(y_n = k | z_{n1}, \ldots, z_{nK}) = \delta(z_{nk} > z_{ni} \ \ \forall i \neq k)$$

Multi-class classification

- The previous treatment can be extended to multiple classes.
- For a problem with K classes:
  - K GP priors, K N-dimensional latent vectors $\mathbf{f}_k$.
  - $N \times K$ auxiliary variables $z_{nk} \sim \mathcal{N}(f_{nk}, 1)$
  - And:

$$P(y_n = k | z_{n1}, \ldots, z_{nK}) = \delta(z_{nk} > z_{ni} \quad \forall i \neq k)$$

- Gibbs sampling is similar to the binary case:
  - Only tricky bit is efficently sampling from a K-dimensionl MVG truncated such that the kth element is largest.

Multi-class classification

- The previous treatment can be extended to multiple classes.
- For a problem with K classes:
  - K GP priors, K N-dimensional latent vectors $\mathbf{f}_k$.
  - $N \times K$ auxiliary variables $z_{nk} \sim \mathcal{N}(f_{nk}, 1)$
  - And:

$$P(y_n = k | z_{n1}, \ldots, z_{nK}) = \delta(z_{nk} > z_{ni} \ \ \forall i \neq k)$$

- Gibbs sampling is similar to the binary case:
  - Only tricky bit is efficently sampling from a K-dimensionl MVG truncated such that the kth element is largest.
- Details of a Variational Bayes inference scheme in: Girolami and Rogers 2006

Multi-class Example

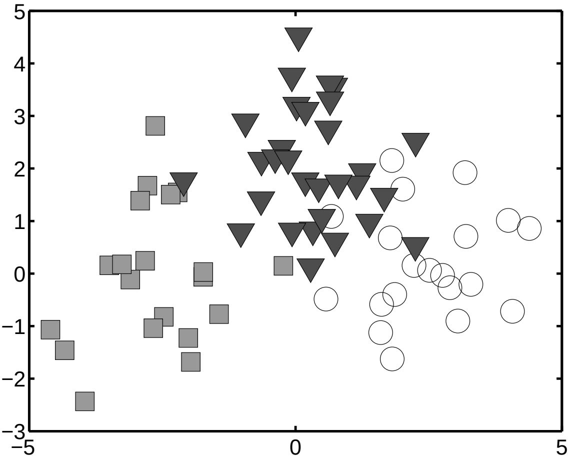


Figure 16 : Multi-class classification example. RBF covariance, $\gamma = 1$.

Multi-class Example

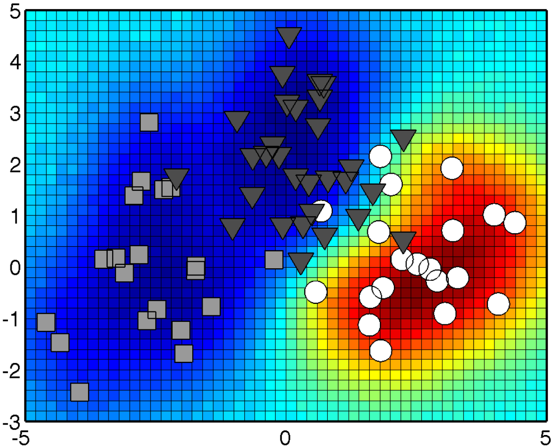


Figure 16 : Multi-class classification example. RBF covariance, $\gamma = 1$.

Multi-class Example

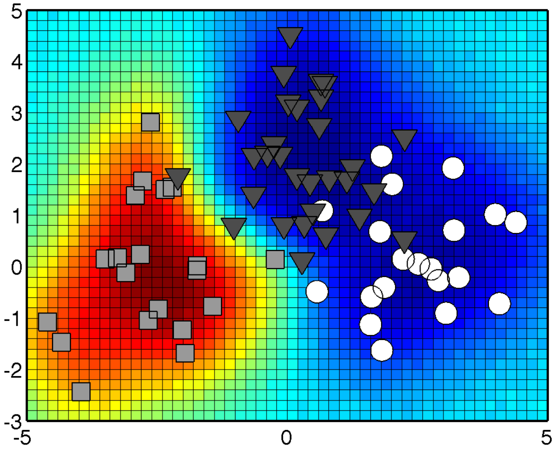


Figure 16 : Multi-class classification example. RBF covariance, $\gamma = 1$.

Multi-class Example

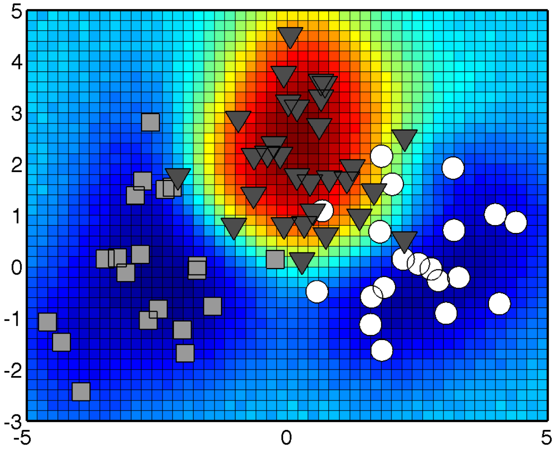


Figure 16 : Multi-class classification example. RBF covariance, $\gamma = 1$.

Semi-supervised Classification

- In some domains, only a subset of data are labeled [e.g. image classification]

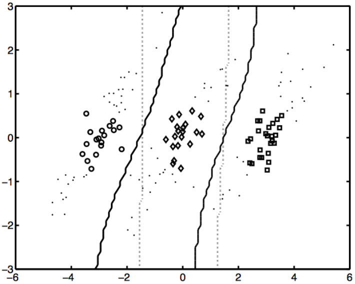


Figure 17 : A toy semi-supervised classification problem.

- Can be overcome using the Null Category Noise Model (NCNM) Lawrence and Jordan 2004

The NCNM

- Going back to binary classification, the auxiliary variable trick can be visualised:

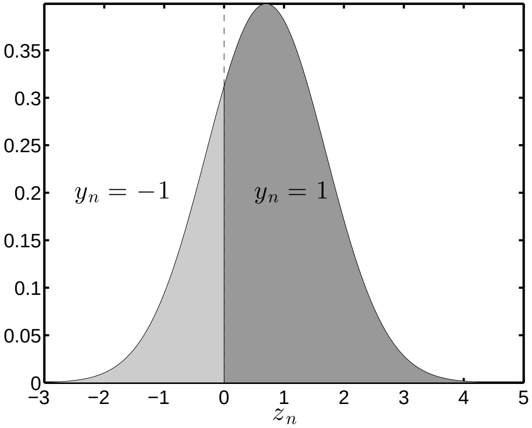


Figure 18 : Visualisation of the auxiliary variable trick. The Gaussian has mean f_n. Note that I'm not calling the classes ± 1.

The NCNM

- To include unlabeled data, we add a third category, for $y_n = 0$:

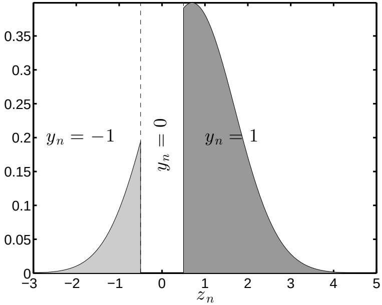


Figure 19 : Visualisation of the NCNM with a null region of width 1.

$$p(y_n|z_n) = \begin{cases} \delta(z_n < -a) & y_n = -1 \\ \delta(z_n > a) & y_n = 1 \\ \delta(z_n > -a) - \delta(z_n > a) & y_n = 0 \end{cases}$$

The NCNM

- The final step is to introduce another set of latent variables.
  - $g_n = 0$ if y_n is observed (i.e. labeled) and $g_n = 1$ otherwise.
- And enforce the constraint that no unlabeled points can exist in the null region:

$$P(y_n = 0 | g_n = 1) = 0$$

The NCNM

- The final step is to introduce another set of latent variables.
  - $g_n = 0$ if y_n is observed (i.e. labeled) and $g_n = 1$ otherwise.
- And enforce the constraint that no unlabeled points can exist in the null region:

$$P(y_n = 0 | g_n = 1) = 0$$

- This has the effect of introducing an empty region around the decision boundary
  - i.e. pushing the decision boundary into regions of empty space

The NCNM

- The final step is to introduce another set of latent variables.
  - $g_n = 0$ if y_n is observed (i.e. labeled) and $g_n = 1$ otherwise.
- And enforce the constraint that no unlabeled points can exist in the null region:

$$P(y_n = 0 | g_n = 1) = 0$$

- This has the effect of introducing an empty region around the decision boundary
  - i.e. pushing the decision boundary into regions of empty space
- Inference:
  - Gibbs sampling is the same as the binary case except $z_n | f_n, g_n = 1$.
  - This is a mixture of two truncated Gaussians – sample the component, and then sample z_n.

NCNM Example

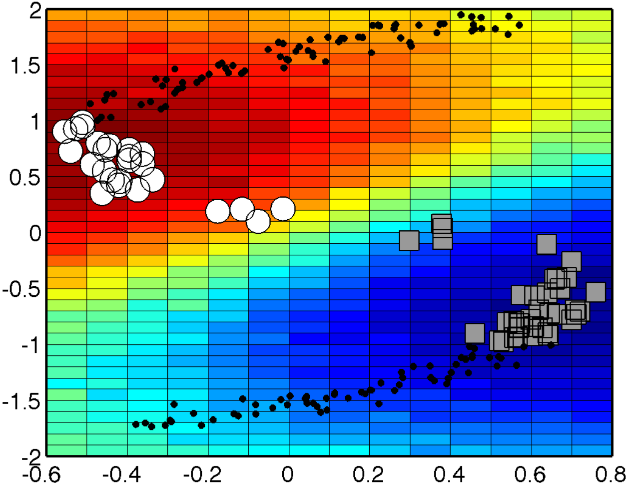


Figure 20 : Standard GP classification (unlabeled data ignored)

NCNM Example

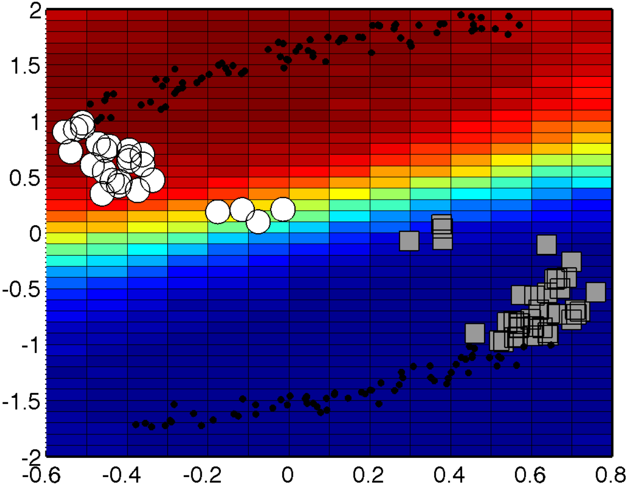


Figure 21 : NCNM GP classification

NCNM Exercise

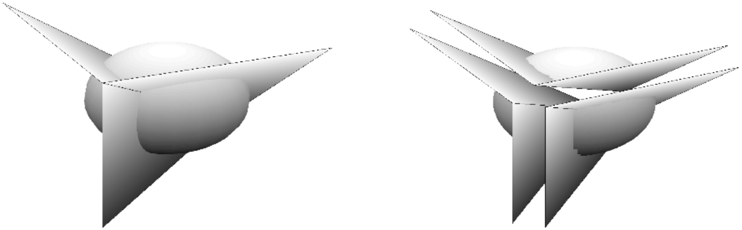
TASK [3]

- Experiment with the NCNM using `gp_ncnm_task.m`
- Setting a=0 results in the standard model
- Setting a>0 uses the NCNM
- It's not always easy to get the results you want to see!

Multi-class NCNM

- This idea can be extended to the multi-class setting.
- See Rogers and Girolami 2007

$$P(y_n = k | z_{n1}, \ldots, z_{nK}) = \begin{cases} \delta(z_{nk} > z_{ni} + a \quad \forall i \neq k) & y_n > 0 \\ 1 - \sum_j \delta(z_{nj} > z_{ni} + a \quad \forall i \neq j) & y_n = 0 \end{cases}$$



(a) A visualisation of the truncation caused by the standard multi-class probit model

(b) A visualisation of the truncation caused by the multi-class probit model with a null region

Figure 22 : Visualisation of truncation

Summary

- GP priors aren't restricted to regression.
- Analytical solutions aren't possible
- Auxiiliary Variable Trick makes inference (via Gibbs sampling or Variational Bayes) straightforward for:
  - Binary classification
  - Ordinal regression
  - Multi-class classification
  - Semi-supervised classification (binary and mutli-class)
  - As well as others (e.g. binary PCA)