Smoothing Spline ANOVA Models III. The Multicategory Support Vector Machine and the Polychotomous Penalized Likelihood Estimate

Grace Wahba

Based on work of Yoonkyung Lee, joint works with Yoonkyung Lee,Yi Lin and Steven A. Ackerman (MSVM) and work of Xiwu Lin (Polychotomous Penalized Likelihood)

`http://www.stat.wisc.edu/˜wahba`
`http://www.stat.ohio-state.edu/˜yklee`
`http://www.stat.wisc.edu/˜yilin`
`http://cimss.ssec.wisc.edu/wxwise/ack.html`
`xiwu_2_lin@spbhrd.com (GlaxoSmithKline)`

[references up on authors’ websites]

Joint Statistical Meetings
San Francisco, CA
August, 2003

Abstract

We describe two modern methods for statistical model building and classification, penalized likelihood methods and and support vector machines (SVM's). Both are obtained as solutions to optimization problems in reproducing kernel Hilbert spaces (RKHS). A training set is given, and an algorithm for classifiying future observations is built from it. The (k-category) multichotomous penalized likelihood method returns a vector of probabilities $(p_1(t), \cdots p_k(t))$ where t is the attribute vector of the object to be classified. The multicategory support vector vachine returns a classifier vector $(f_1(t), \cdots f_k(t))$ satisfying $\sum_\ell f_\ell(t) = 0$, where $max_\ell f_\ell(t)$ identifies the category. The two category SVM's are very well known, while the multi-category SVM (MSVM) described here, which includes modifications for unequal misclassification costs and unrepresentative training sets, is new.

We describe applications of each method: For penalized likelihood, estimating the 10-year probability of death due to several causes, as a function of several risk factors observed in a demographic study, and for MSVM's, classifying radiance profiles from the MODIS instrument according to clear, water cloud or ice cloud. Some computational and tuning issues are noted.

OUTLINE

0. Review of the regularization problem in RKHS.

1. Optimal classification and the Neyman-Pearson Lemma.

2. Penalized likelihood estimation, two classes.

3. The Support Vector Machine, two classes.

4. Penalized likelihood and the SVM compared.

5. Multichotomous penalized likelihood estimates.

6. Multicategory support vector machines (MSVMs)

7. Tuning the estimates.

8. Application to cloud classification from MODIS data.

9. Closing remarks.

References

V. Vapnik *et al*.on SVM's, www.kernel-machines.org.

[Lee02] Y. Lee *Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data* PhD. thesis, also UW-Madison TR 1063.

[LeeLinWahba02] Y. Lee, Y. Lin and G. Wahba. Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data, TR 1064, subm .

[YLin02] Y. Lin. Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275, 2002.

[LWA03] Y. Lee, G. Wahba and S. Ackerman. Classification of Satellite Radiance Data by Multicategory Support Vector Machines, TR 1075, subm.

[XLin98] X. Lin. *Smoothing Spline Analysis of Variance for Polychotomous Response Data*. PhD thesis, Department of Statistics, Uiversity of Wisconsin, Madison WI, 1998. Also TR 1003, available via G. Wahba home page.

[Wahba02]G. Wahba. Soft and hard classification by reproducing kernel Hilbert space methods. Proc. NAS 2002, 99, 16524-16503.

[Wahba99] G. Wahba. Support Vector Machines, Reproducing Kernel Hilbert Spaces and the Randomized GACV. In ‘Advances in Kernel Methods - Support Vector Learning’, Schölkopf, Burges and Smola (eds.), MIT Press 1999, 69-88.

[XiangWahba96]D. Xiang, D. and G. Wahba. A Generalized Approximate Cross Validation for Smoothing Splines with Non-Gaussian Data, Statistica Sinica, 6, 1996, pp.675-692.

♣♣ 0. Regularization Problems in RKHS from Lecture 1

To set notation: The canonical regularization problem in RKHS we discussed in the first lecture was: Given

$$\{y_i, t_i\}, y_i \in \mathcal{S}, t_i \in \mathcal{T},$$

and $\{\phi_1, \cdots, \phi_M\}$, M special functions defined on $\mathcal{T}$, find f of the form

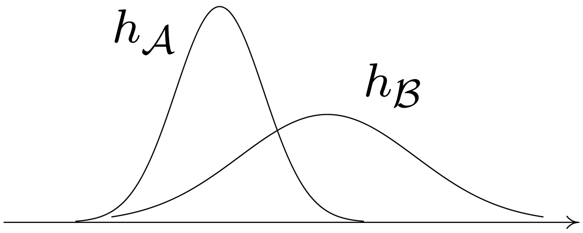
$$f = \sum_{\nu=1}^{M} d_\nu \phi_\nu + h$$

with $h \in \mathcal{H}_K$ to minimize

$$\mathcal{I}_\lambda\{f, y\} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i)) + \lambda \|h\|^2_{\mathcal{H}_K}.$$

Today y_i will just be a class label, $y_i \in \mathcal{S} = \{1, 2, \cdots, k\}$ (k classes)

♣♣ 1. Optimal Classification and the Neyman-Pearson Lemma:



$h_{\mathcal{A}}(\cdot), h_{\mathcal{B}}(\cdot)$ densities of t for class $\mathcal{A}$ and class $\mathcal{B}$.

NOTATION:

$\pi_{\mathcal{A}} =$ prob. next observation (Y) is an $\mathcal{A}$

$\pi_{\mathcal{B}} = 1 - \pi_{\mathcal{A}} =$ prob. next observation is a $\mathcal{B}$

$$\begin{aligned} p(t) &= prob\{Y = \mathcal{A}|t\} \\ &= \frac{\pi_{\mathcal{A}} h_{\mathcal{A}}(t)}{\pi_{\mathcal{A}} h_{\mathcal{A}}(t) + \pi_{\mathcal{B}} h_{\mathcal{B}}(t)} \end{aligned}$$

♣♣ 1.Optimal Classification and the Neyman-Pearson Lemma (cont.).

Let $c_{\mathcal{A}} =$ cost to falsely call a $\mathcal{B}$ an $\mathcal{A}$

$c_{\mathcal{B}} =$ cost to falsely call an $\mathcal{A}$ a $\mathcal{B}$

Bayes classification rule: Let

$\phi(t)$: $\quad t \rightarrow \{{\mathcal{A} \atop \mathcal{B}}\}$

Optimum (Bayes) classifier: (Neyman-Pearson Lemma) Minimizes the expected cost:

$$\phi_{\text{OPT}}(t) = \begin{cases} \mathcal{A} & \text{if } \frac{p(t)}{1-p(t)} > \frac{c_{\mathcal{A}}}{c_{\mathcal{B}}}, \\ \mathcal{B} & \text{otherwise.} \end{cases}$$

♣♣ 2. Penalized likelihood estimation, two classes.

Let $f(t) = \log p(t)/(1 - p(t))$, the log odds ratio. Assume (for simplicity only) $\frac{c_{\mathcal{A}}}{c_{\mathcal{B}}} = 1$

Then the optimal classifier is

$$f(t) > 0 \text{ (equivalently, } p(t) - \tfrac{1}{2} > 0) \rightarrow \mathcal{A}$$
$$f(t) < 0 \text{ (equivalently, } p(t) - \tfrac{1}{2} < 0) \rightarrow \mathcal{B}$$

To estimate f: Assume (again for simplicity only) that the relative frequency of $\mathcal{A}$'s in the training set is about the same as in the general population:

♣♣ 2. Penalized likelihood estimation,two classes(cont.).

Code the data as:

$$y_i = \begin{matrix} 1 & = \mathcal{A} \\ 0 & = \mathcal{B} \end{matrix} \text{ (important)}$$

The probability distribution function (likelihood) for $y \mid p$:

$$\mathcal{L}(y,p) = p^y(1-p)^{1-y} = \begin{cases} p & \text{if } y = 1 \\ (1-p) & \text{if } y = 0 \end{cases}$$

and the negative log likelihood is

$$\begin{aligned} -\log \mathcal{L} &= -\log[p^y(1-p)^{1-y}] \\ &= -y\log p - (1-y)\log(1-p). \end{aligned}$$

Substituting $p = e^f/(1+e^f)$ gives

$$-\log \mathcal{L}(y,f) = -yf + \log(1+e^f).$$

The penalized log likelihood estimate of f is obtained by setting

$$\mathcal{C}(y_i, f(t_i)) = -y_i f(t_i) + \log(1 + e^{f(t_i)})$$

in the optimization problem $\mathcal{I}_\lambda(f,y)$.

♣♣ 3. The Support Vector Machine, two classes.

$$y = \begin{matrix} +1 = \mathcal{A} \\ -1 = \mathcal{B} \end{matrix} \text{ (note different coding)}$$

Find $f(t) = d + h(t)$ with $h \in \mathcal{H}_K$ to min

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(t_i))_+ + \lambda \|h\|^2_{\mathcal{H}_K} \qquad (**)$$

where $(\tau)_+ = \tau, \tau > 0, = 0$ otherwise.

Then

$$f_\lambda(t) = d + \sum_{i=1}^{n} c_i K(t, t_i). \qquad (*)$$

Substitute (*) into (**), choose λ, given λ, find c and d. The classifier is

$$f_\lambda(t) > 0 \rightarrow \mathcal{A}$$

$$f_\lambda(t) < 0 \rightarrow \mathcal{B}$$

♣♣ 4. Penalized likelihood estimation and the SVM compared:

Let us relabel y in the likelihood –

$$\tilde{y}_i = \begin{cases} +1 & \text{if } \mathcal{A}, \\ -1 & \text{if } \mathcal{B}. \end{cases}$$

Then

$$-yf + \log(1 + e^{f}) \rightarrow \log(1 + e^{-\tilde{y}f})$$

Figure 1 compares

$$\log(1 + e^{-yf}), \quad (1 - yf)_{+} \text{ and } (-yf)_{*}$$

where

$$(\tau)_{*} = \begin{cases} 1 & \text{if } \tau > 0, \\ 0 & \text{otherwise.} \end{cases}$$

$(-yf)_{*}$ is the misclassification counter.
$(1 - yf)_{+}$ is known as the ”hinge function”.

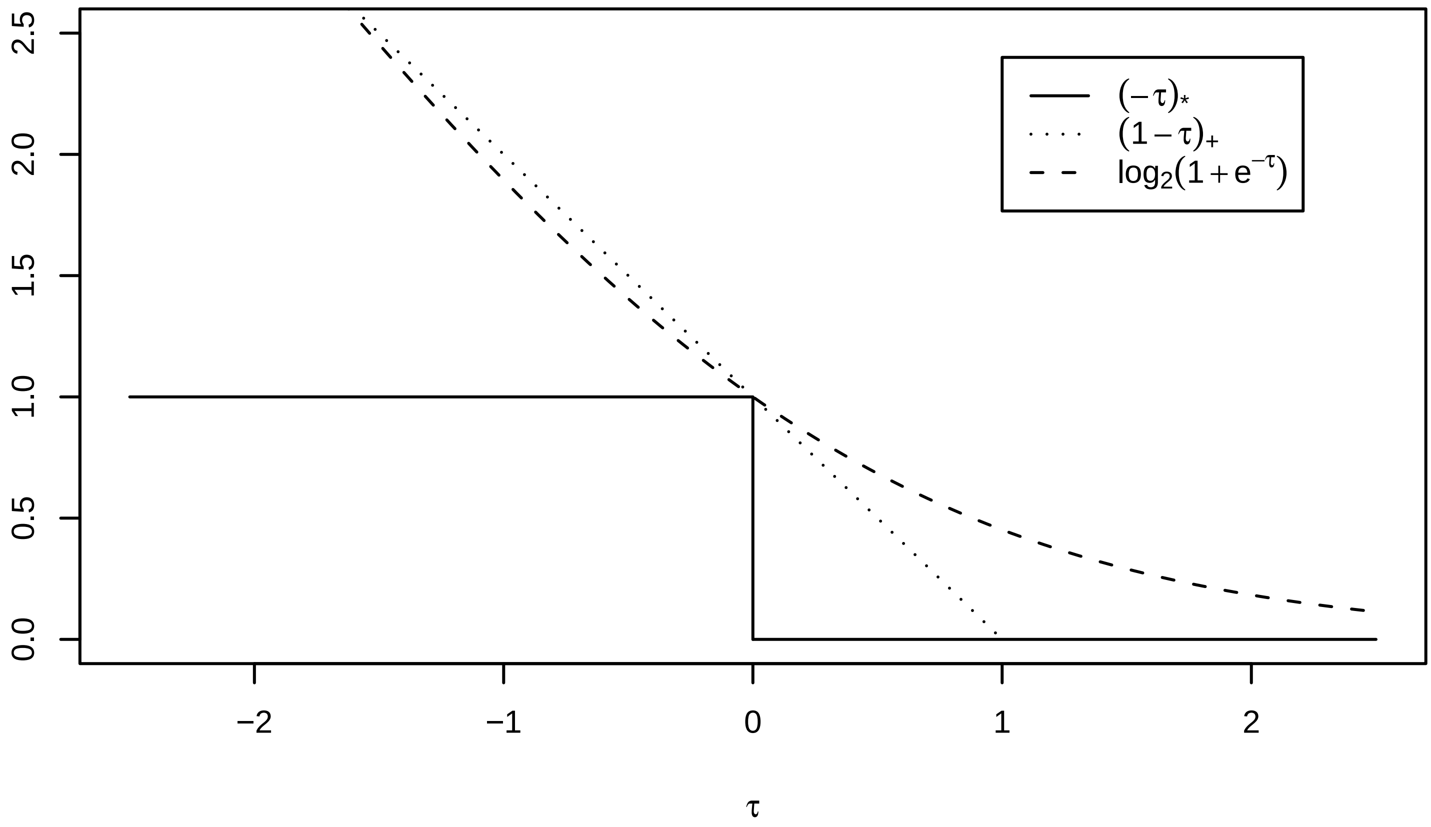


Figure 1. Let $\mathcal{C}(y_i, f(t_i)) = c(y_i f(t_i)) = c(\tau)$. Comparison of $c(\tau) = (-\tau)_*, (1-\tau)_+$ and $log_2(1+e^{-\tau})$. Any strictly convex function that goes through 1 at $\tau = 0$ will be an upper bound on the missclassification counter $(-\tau_*)$ and will be a looser bound than some SVM (hinge) function $(1 - \theta\tau)_+$. Many other ”large margin” classifiers. (See [Wahba02]).

♣♣ 4. Penalized likelihood and the SVM compared(cont.).

The penalized log likelihood estimate is tuned by a criteria which chooses λ to minimize a proxy for

$$R(\lambda) = E\frac{1}{n}\sum_{i=1}^{n} -y_{new\cdot i}f_\lambda(t_i) + \log(1 + e^{f_\lambda(t_i)}).$$

[XiangWahba96]. $R(\lambda)$ is the expected ‘distance’ or negative log likelihood for a new observation with the same t_i. $f_{\lambda_{opt}}$ estimates the log odds ratio $log[p/(1-p)]$.

We say the SVM classifier is optimally tuned if we have a criteria which chooses λ to minimize a proxy for

$$R(\lambda) = E\frac{1}{n}\sum_{i=1}^{n}(1 - y_{new\cdot i}f_\lambda(x_i))_+.$$

That is, it is choosing λ to minimize a proxy for an an upper bound on the misclassification rate [LeeLin-Wahba02][Wahba99].

♣♣ 4. Penalized likelihood and the SVM compared(cont.).

What is the SVM estimating?

Lemma (Yi Lin [YLin02])

The minimizer of $E(1 - y_{new} f(t))_+$ is $sign\ f(t)$ $(= sign\ (p(t) - \frac{1}{2}) = sign\ (2p(t) - 1))$

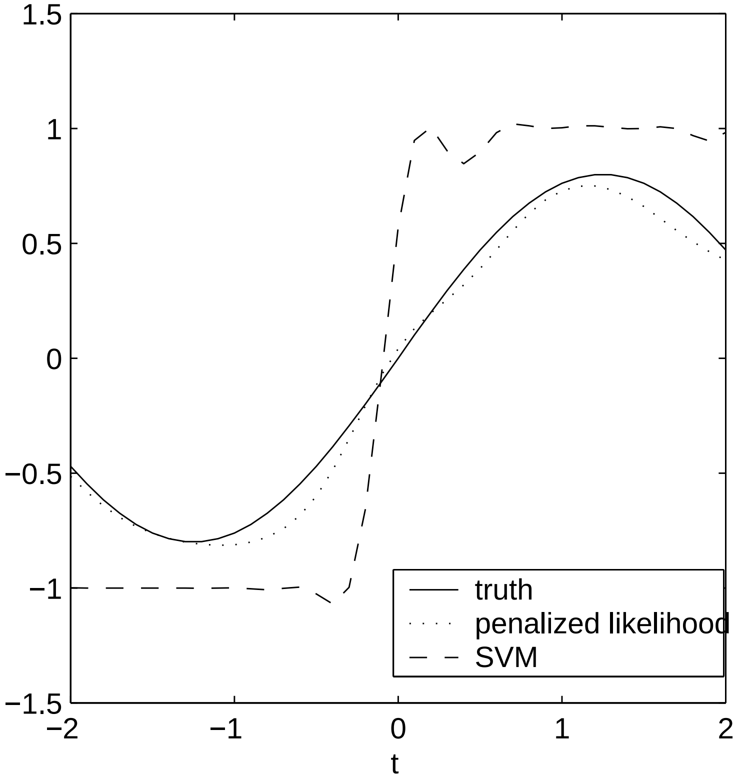
where $f(t) = \log p(t)/(1 - p(t))$.

So the SVM, the solution of the problem: Find $f_\lambda = d + h$ which minimizes

$$\frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(t_i))_+ + \lambda \|h\|^2_{\mathcal{H}_K},$$

where λ is chosen to minimize (a proxy for) $R(\lambda)$, is estimating sign $f(t)$ - not $f(t)$ itself, but just what you need to minimize the misclassification rate.

♣♣ 4. Penalized likelihood and the SVM compared(cont.).



300 Bernoulli random variables were generated, equally spaced t from $p(t) = 0.4sin(0.4\pi t) + 0.5$ Solid line: $(2p(t) - 1)$. Dotted line:$(2p_\lambda - 1)$, p_λ is (optimally tuned) penalized likelihood estimate of p. Dashed line: $f_{svm\ \lambda}$, is (optimally tuned) SVM. Observe $f_{svm\ \lambda} \sim \pm 1$, thus p_λ is estimating $p(t)$, whereas $f_{svm\ \lambda}$ is estimating $sign(2p-1)$ = $sign(p-1/2)$= $sign\ f$. (based on Gaussian K) (plot: Yoonkyung Lee)

♣♣ 5. Multichotomous penalized likelihood[XLin98].

$k+1$ categories, $k>1$. Let $p_j(t)$ be the probability that a subject with attribute vector t is in category j, $\sum_{j=0}^{k} p_j(t)=1$. From [XLin98]: Let

$$f^j(t)=\log p_j(t)/p_0(t), j=1, \cdots, k.$$

Then:

$$\begin{aligned} p_j(t) &= \frac{e^{f^j(t)}}{1+\sum_{j=1}^{k} e^{f^j(t)}},\ j=1,\cdots,k \\ p_0(t) &= \frac{1}{1+\sum_{j=1}^{k} e^{f^j(t)}} \end{aligned}$$

Coding:

$$y_i=(y_{i1},\cdots,y_{ik}),$$

$y_{ij}=1$ if the ith subject is in category j and 0 otherwise.

♣♣ 5. Multichotomous penalized likelihood (cont.).

Letting $f = (f^1, \cdots, f^k)$ the negative log likelihood can be written as $-log\mathcal{L}(y, f)$

$$= \sum_{i=1}^{n} \{- \sum_{j=1}^{k} y_{ij} f^j(t_i) + log(\sum_{j=1}^{k} 1 + e^{f^j(t_i)})\}.$$

where

$$f^j = \sum_{\nu_j=1}^{M} d_{\nu j} \phi_\nu + h^j.$$

$\lambda \|h\|^2_{\mathcal{H}_K}$ becomes

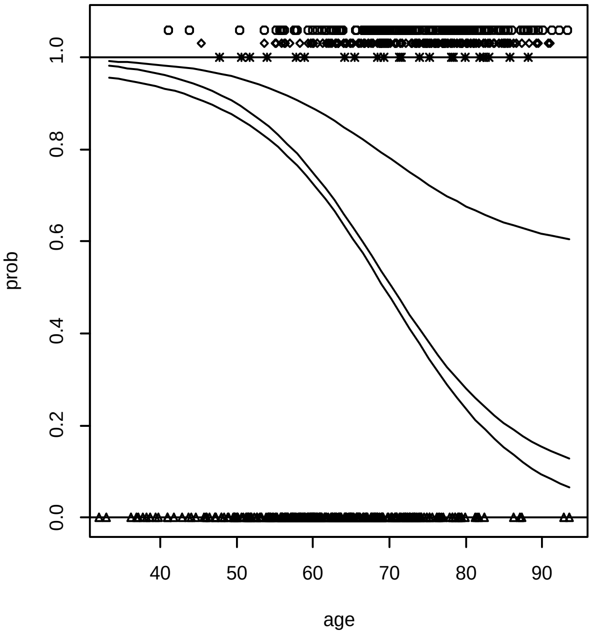
$$\sum_{j=1}^{k} \lambda_j \|h^j\|^2_{\mathcal{H}_K},$$

and the optimization problem becomes: Minimize

$$I_\lambda(y, f) = -log\mathcal{L}(y, f) + \sum_{j=1}^{k} \lambda_j \|h^j\|^2_{\mathcal{H}_K}.$$

♣♣ 5. Multichotomous penalized likelihood (cont.).

10 year risk of mortality as a function of $t = (x_1, x_2, x_3) =$ age, glycosylated hemoglobin, and systolic blood pressure[From XLin98].



x_2 and x_3 set at their medians. The differences between adjacent curves (from bottom to top) are probabilities $p_j(t)$ for : 0:alive, 1: diabetes, 2: heart attack, 3: other causes. $f^j(x_1, x_2, x_3) =$

$$\mu^j + f_1^j(x_1) + f_2^j(x_2) + f_3^j(x_3) + f_{23}^j(x_2, x_3)$$

(Smoothing Spline ANOVA model.)

♣♣ 6. Multicategory support vector machines (MSVMs).

From [LeeLinWahba02],[LWA03], earlier reports.
$k > 2$ categories. Coding:

$$y_i = (y_{i1}, \cdots, y_{ik}), \sum_{j=1}^{k} y_{ij} = 0,$$

in particular $y_{ij} = 1$ if the ith subject is in category j and $y_{ij} = -\frac{1}{k-1}$ otherwise. $y_i = (1, -\frac{1}{k-1}, \cdots, -\frac{1}{k-1})$ indicates y_i is from category 1. The MSVM produces $f(t) = (f^1(t), \cdots f^k(t))$, with each $f^j = d^j + h^j$ with $h^j \in \mathcal{H}_K$, *required to satisfy a sum-to-zero constraint*

$$\sum_{j=1}^{k} f^j(t) = 0,$$

for all t in $\mathcal{T}$. The largest component of f indicates the classification.

♣♣ 6. Multicategory support vector machines (MSVMs)(cont.).

Let $L_{jr} = 1$ for $j \neq r$ and 0 otherwise. The MSVM is defined as the vector of functions $f_\lambda = (f_\lambda^1, \cdots, f_\lambda^k)$, with each h^k in $\mathcal{H}_K$ satisfying the sum-to-zero constraint, which minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{k} L_{cat(i)r}(f^r(t_i) - y_{ir})_+ + \lambda \sum_{j=1}^{k} \|h^j\|^2_{\mathcal{H}_K}$$

equivalently

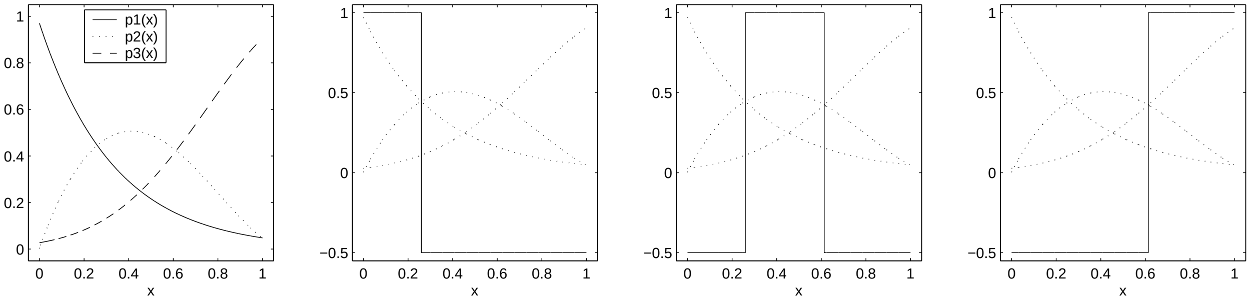
$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r\neq cat(i)} (f^r(t_i) + \frac{1}{k-1})_+ + \lambda \sum_{j=1}^{k} \|h^j\|^2_{\mathcal{H}_K}$$

where $cat(i)$ is the category of y_i.

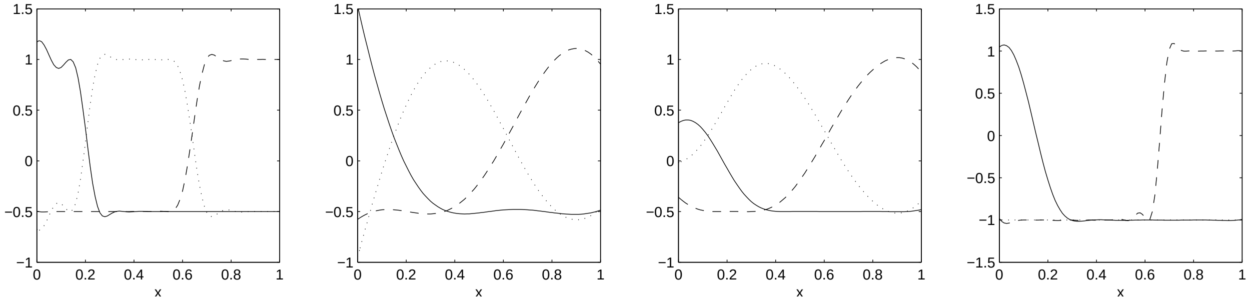
The $k = 2$ case reduces to the usual 2-category SVM.

The target for the MSVM is $f(t) = (f^1(t), \cdots, f^k(t))$ with $f^j(t) = 1$ if $p_j(t)$ is bigger than the other $p_l(t)$ and $f^j(t) = -\frac{1}{k-1}$ otherwise.

♣♣ 6. Multicategory support vector machines (MSVMs)(cont.).



Above: Probabilities and target f^j's for three category SVM demonstration.(Gaussian Kernel)



The left panel above gives the estimated f^1, f^2 and f^3. λ and σ were optimally tuned. (i. e. with the knowledge of the 'right' answer). In the second from left panel both λ and σ were chosen by 5-fold cross validation in the MSVM and in the third panel they were chosen by GACV. In the rightmost panel the classification is carried out by a one-vs-rest method.

♣♣ 6. Multicategory support vector machines(MSVMs)(cont.).

The nonstandard MSVM:

More generally, suppose the sample is not representative, and misclassification costs are not equal. Let

$$L_{jr} = (\pi_j / \pi_j^s) C_{jr}, \quad j \neq r$$

C_{jr} is the cost of misclassifying a j as an r, $C_{rr} = 0$, π_j is the prior probability of category j, and π_j^s is the fraction of samples from category j in the training set. Then the nonstandard MSVM has as its target the Bayes rule, which is to choose the j which minimizes

$$\sum_{\ell=1}^{k} C_{\ell j} p_\ell(x)$$

♣♣ 7. Tuning the estimates.

GACV (generalized approximate cross validation). Penalized likelihood: [XiangWahba96][XLin98]; SVM[Wahba99], MSVM[Lee02][LeeLinWahba02].

Leaving out one:

$$V_O(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{C}(y_i, f_\lambda^{[i]}(t_i))$$

where $f_\lambda^{[i]}$ is the estimate without the ith data point.

$$GACV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i)) + D(y, f_\lambda)$$

where

$$D(y, f_\lambda) \approx \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathcal{C}(y_i, f_\lambda^{[i]}(t_i)) - \mathcal{C}(y_i, f_\lambda(t_i)) \right\}$$

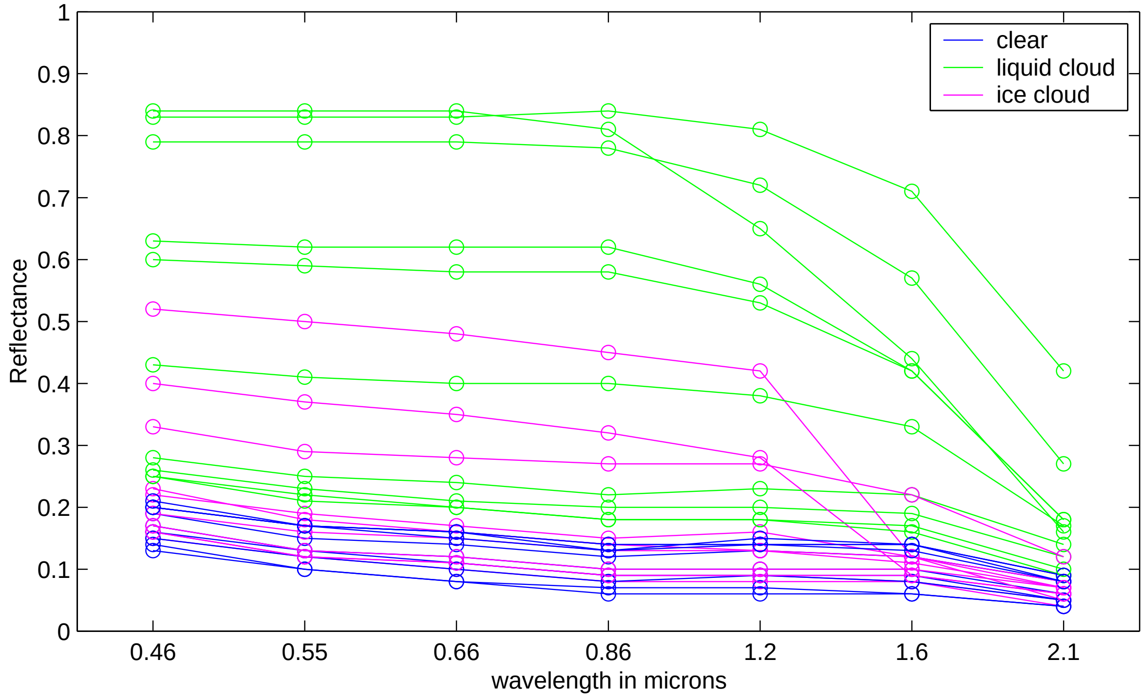
is obtained by a tailored perturbation argument. Easy to compute for the SVM, use randomized trace techniques to estimate the perturbation in the likelihood case.

♣♣ 8. The classification of upwelling MODIS radiance data to clear sky, water clouds or ice clouds.

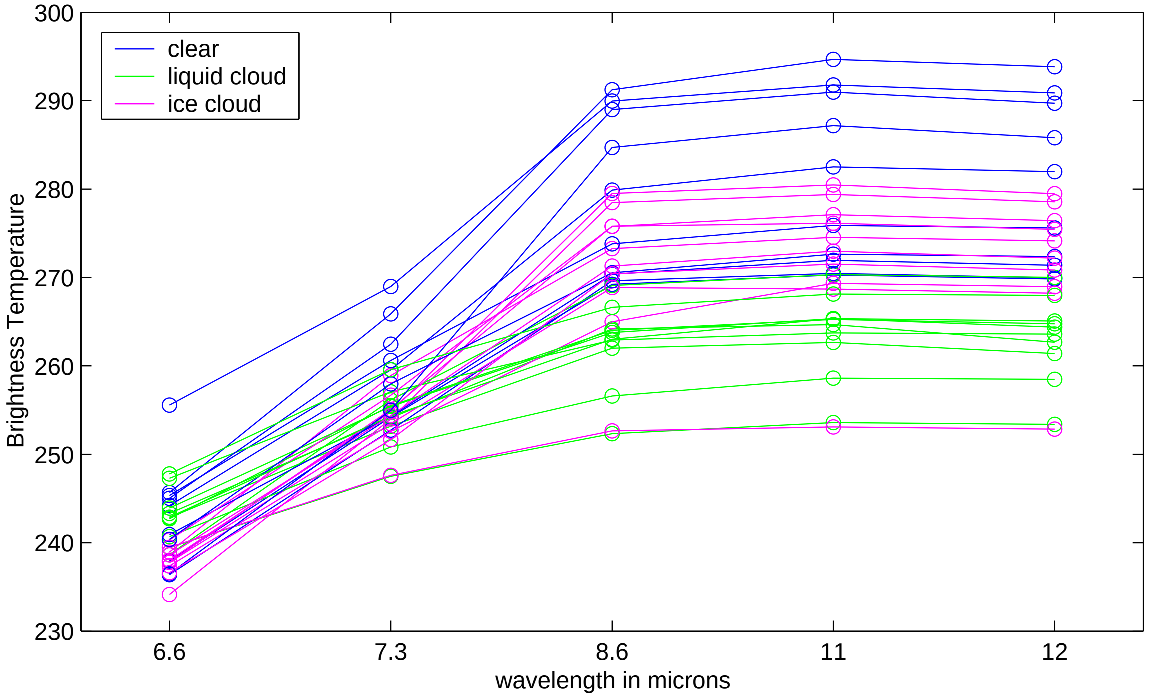
From [LWA03].Classification of 12 channels of upwelling radiance data from the satellite- borne MODIS instrument. MODIS is a key part of the Earth Observing System (EOS).

Classify each vertical profile as coming from clear sky, water clouds, or ice clouds.

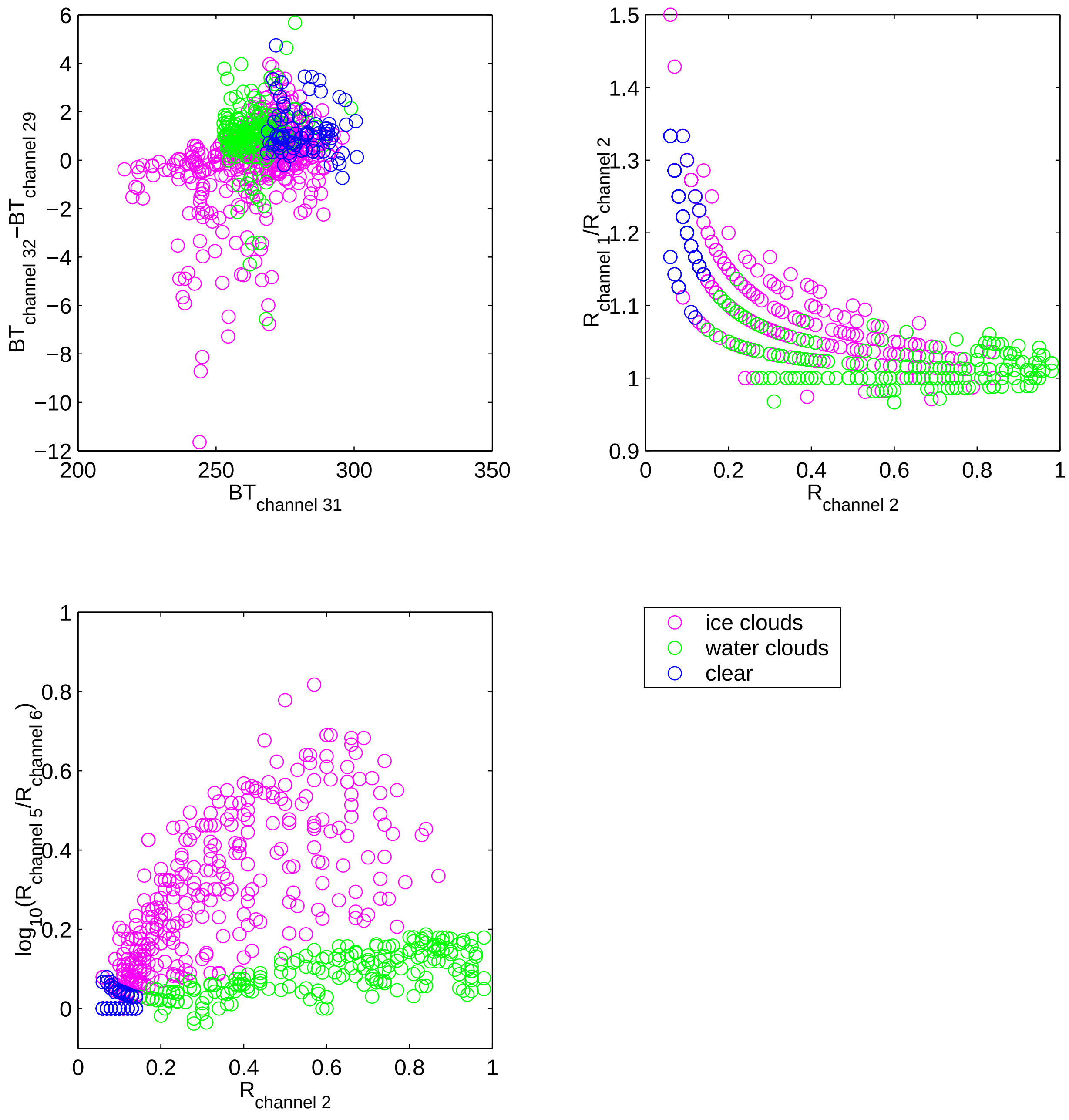
Next page: 744 simulated radiance profiles (81 clear-blue, 202 water clouds-green, 461 ice clouds-purple). 10 samples from clear, from water and from ice:



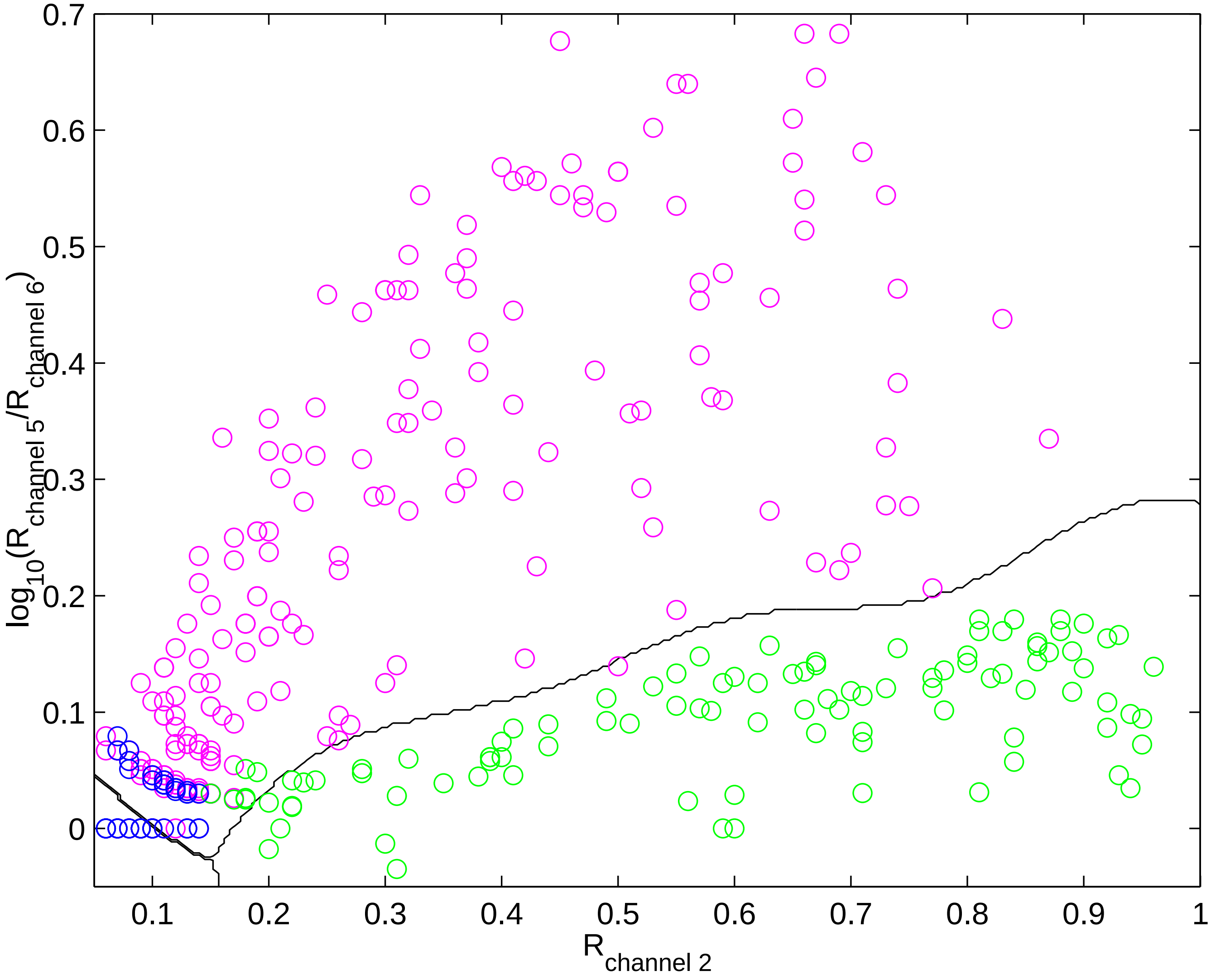
clear
liquid cloud
ice cloud
Reflectance
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.46
0.55
0.66
0.86
1.2
1.6
2.1
wavelength in microns



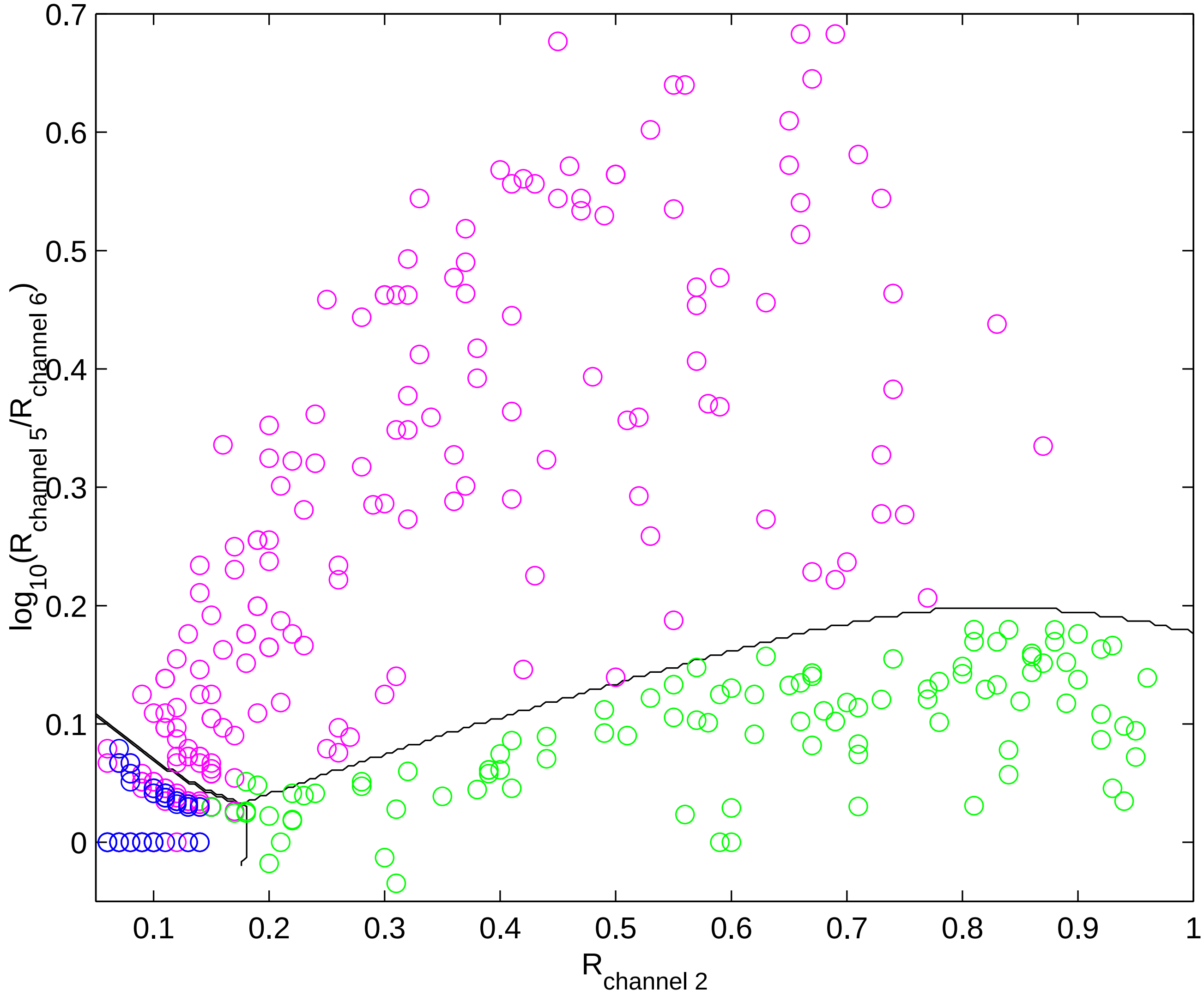
clear
liquid cloud
ice cloud
Brightness Temperature
300
290
280
270
260
250
240
230
6.6
7.3
8.6
11
12
wavelength in microns



Pairwise plots of three different variables (including composite variables.(purple = ice clouds, green = water clouds, blue = clear)



Classification boundaries on the 374 test set determined by the MSVM using 370 training examples, two variables, one is composite. Y. K. Lee Student poster prize AMetSoc Satellite Meteorology and Oceanography session.



Classification boundaries determined by the nonstandard MSVM when the cost of misclassifying clear clouds is 4 times higher than other types of misclassifications.

♣♣ 9. Closing Remarks

We have examined two class and multi-class penalized likelihood estimates and Support Vector Machines, both of which can be obtained as optimization problems in an RKHS. Non-representative samples and unequal misclassification costs can be handled. Newton-Raphson and Mathematical Programming are used to solve the optimization problems. Downhill simplex works well for searching multiple λ's. Convergence theory of various penalized likelihood models have been around a long time, convergence theory for the SVM (to sign f) and MSVM (to its target) is younger. Penalized likelihood estimates and SVM's are just two of the many examples of optimization problems related to regularization in RKHS, which have many useful scientific applications.