The Multicategory Support Vector Machine, with Application to the Classification of Simulated and Real MODIS Data into Clear, Ice Cloud and Water Cloud Categories.

Grace Wahba

Based on Lee, Wahba and Ackerman, to appear JTECH available via my home page or the prepublication directory of JTECH. Yoonkyung Lee won the "Best Student Poster" award at the AMetSoc 2003 Satellite and Oceanography Session.

`http://www.stat.wisc.edu/~wahba`
`http://www.stat.ohio-state.edu/~yklee`
`http://cimss.ssec.wisc.edu/wxwise/ack.html`

[references up on authors' websites]

Atmospheric and Oceanic Sciences Colloquium
Madison WI
December 8, 2003

Abstract

We describe a modern method for statistical classification known as support vector machine (SVM). The two class SVM has been known for a decade. The multicategory SVM (MSVM) was originally proposed in the thesis of Yoonkyung Lee (2002) and joint technical reports by various subsets of Lee, Lee, Lin, Wahba and Zhang. Recently Lee and Wahba teamed up with Steve Ackerman to apply the results to simulated MODIS data, to classify the MODIS profiles as coming from a clear, water cloud or ice cloud situation. Very good results were obtained. When the JTECH referee wanted to know how well the method would work on real MODIS data, a satellite ”expert” took a sample of 1536 observed MODIS profiles, and labeled them with other information at his disposal. The labeled profiles were then divided into a training set and a test set. The MSVM built on the training set achieved an error rate of under 1% on the test set, while the present MODIS algorithm had an error rate of about 18%. As ann interesting byproduct we note how closely the simulated data matched the observational data in some of pairwise variable plots. (To appear, JTECH).

OUTLINE

0. Teaser: Simulated MODIS data

1. Optimal classification and the Neyman-Pearson Lemma.

2. The Support Vector Machine (SVM), two classes.

3. Tuning the estimates

4. The Multicategory Support Vector Machine (MSVM)

5. Application to cloud classification from MODIS data.

6. Closing remarks.

References

The Support Vector Machine Website: www.kernel-machines.org.

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

poster prizewinner

[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, to appear JASA.

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

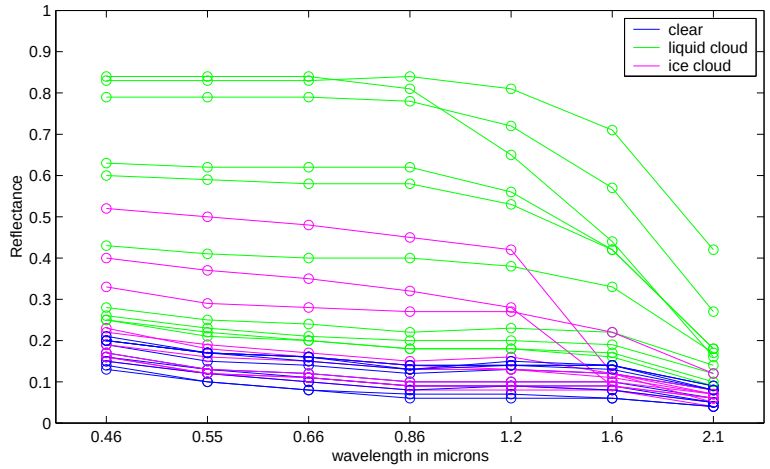
[Wahba02]G. Wahba. Soft and hard classification by reproducing kernel Hilbert space methods. Proc. NAS 2002, 99, 16524-16503.

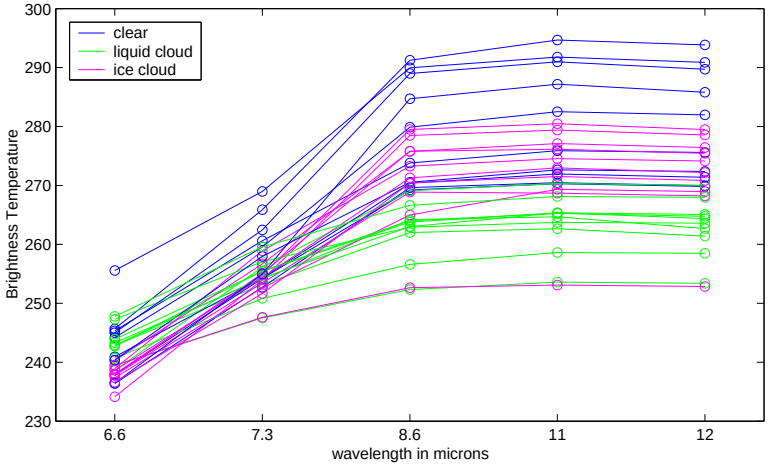
[Wahba99] G. Wahba. Support Vector Machines, Reproducing Kernel Hilbert Spaces and the Randomized GACV. In 'Advances in Kernel Methods - Support Vector Learning', Schölkopf, Burges and Smola (eds.), MIT Press 1999, 69-88.

[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.

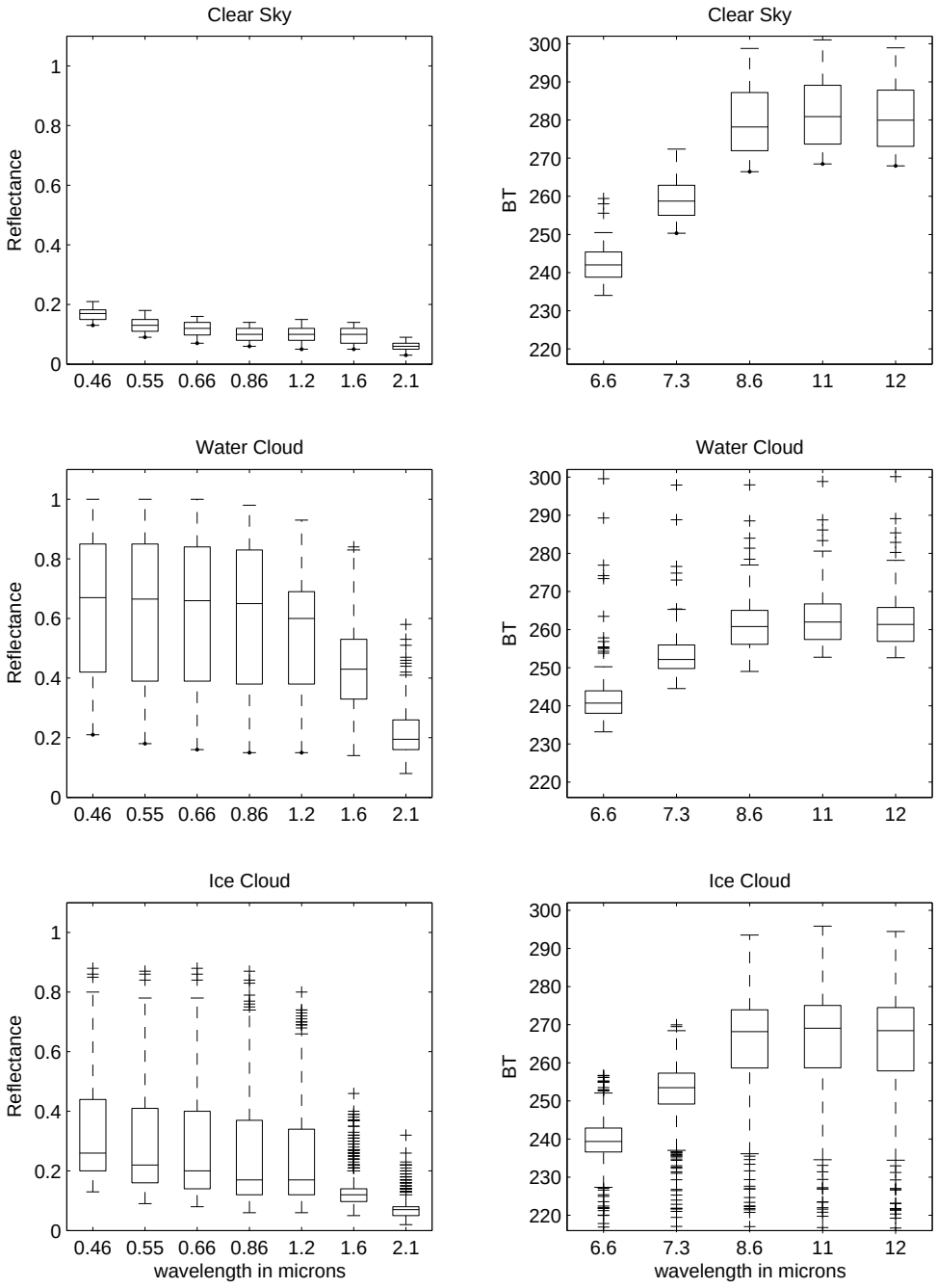
♣♣ 0. Simulated MODIS Data

744 radiance profiles were simulated (81 clear, 202 water clouds, 461 ice clouds). Here are 10 samples from clear, from water clouds and from ice clouds:

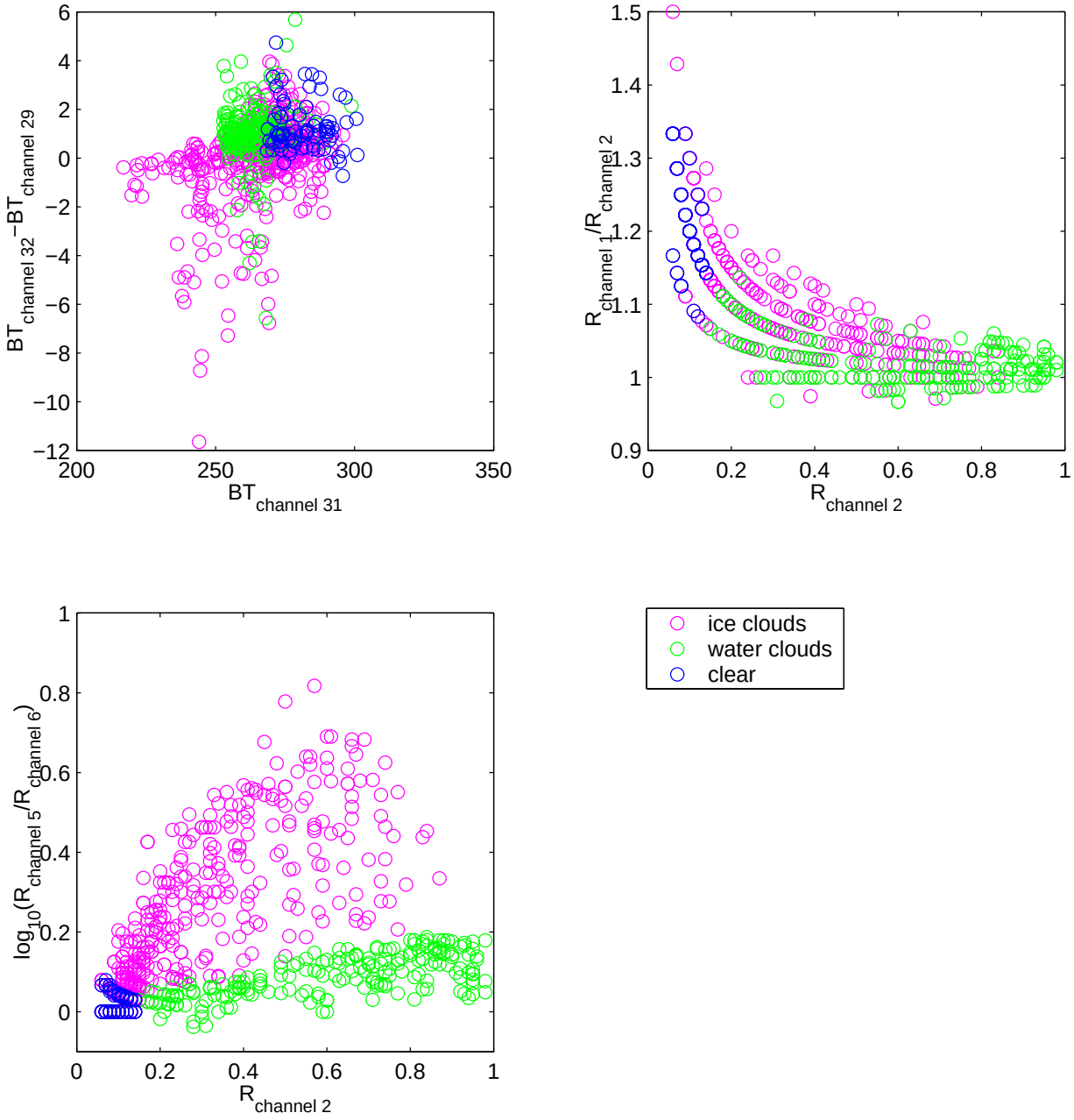




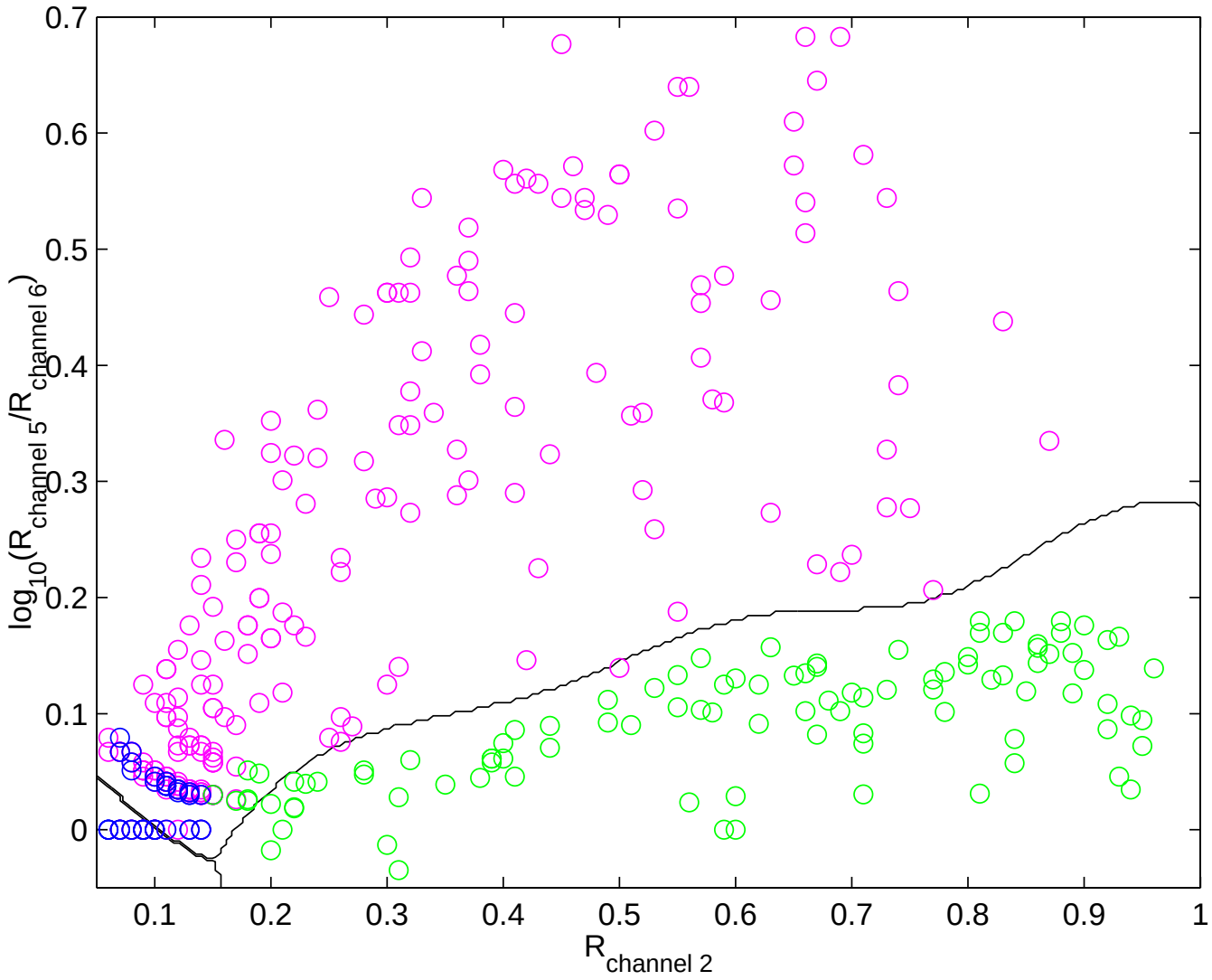
(purple = ice, green = water, blue = clear)



Boxplots of 7 reflectances and 5 brightness temperatures, clear, water clouds, ice clouds (over ocean).

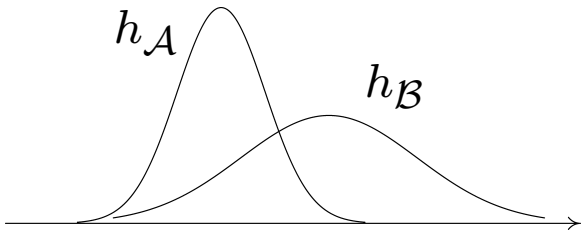


Pairwise plots of three different variables (including composite variables.(purple = ice clouds, green = water clouds, blue = clear)



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.

♣♣ 1. Optimal Classifi cation 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

$\phi(t): \quad t \rightarrow \{\begin{smallmatrix}\mathcal{A}\\ \mathcal{B}\end{smallmatrix}\}$

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

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

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

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

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

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

$$\|h\|^2_{\mathcal{H}_K} = \sum_{i,j} c_i c_j K(t_i, t_j). \qquad (**)$$

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

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

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

Numerically, must solve a mathematical programming problem.

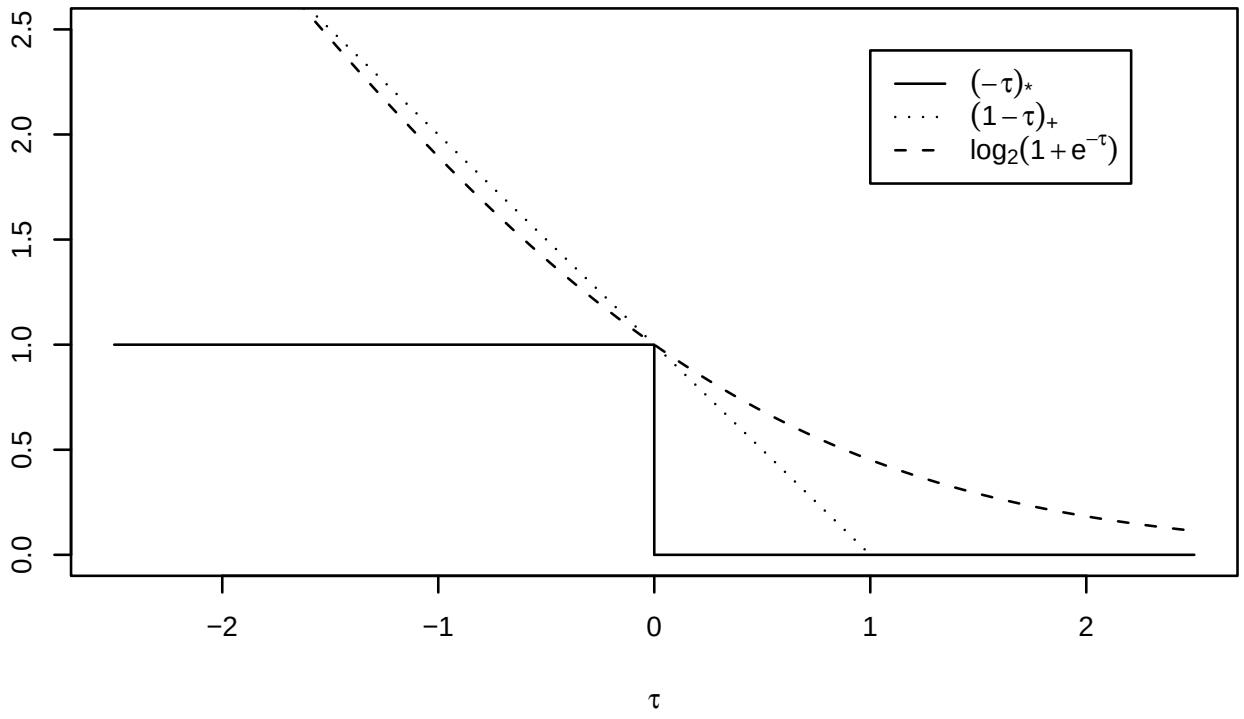


Figure 1. Let $\mathcal{C}(y_i, f(t_i)) = c(y_i f(t_i)) = c(\tau)$. Comparison of $c(\tau) = (-\tau)_*$, $(1-\tau)_+$ and $log_2(1+e^{-\tau})$, the log likelihood function. Any strictly convex function that goes through 1 at $\tau = 0$ will be an upper bound on the missclassification counter $(-\tau_*)$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

♣♣ 2.The SVM (cont.) What is the SVM estimating?.

What is the SVM estimating?

Lemma (Yi Lin [YLin02]) (two category version)

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

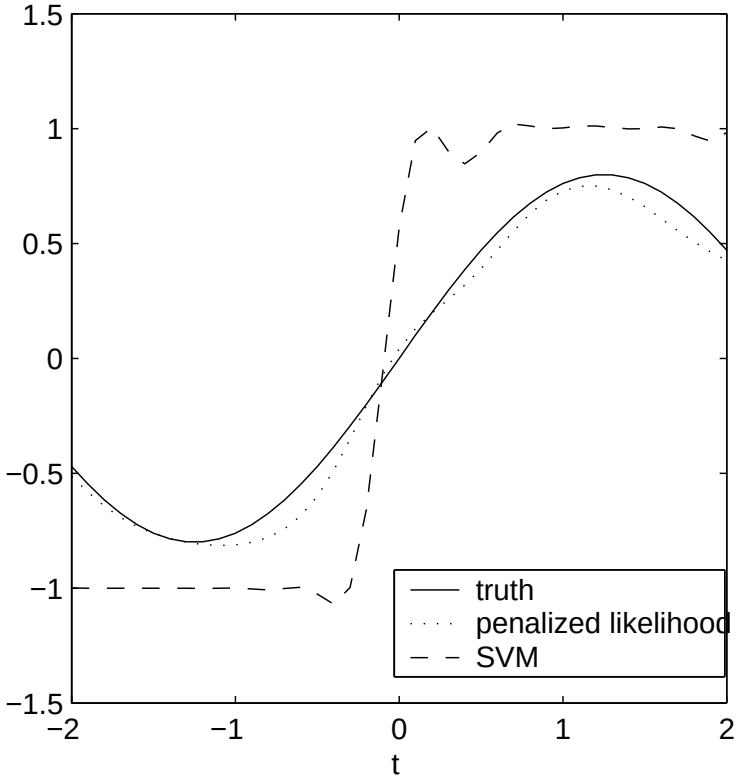
where $f(t) = \log p(t)/(1 - p(t))$.

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.

♣♣ 2. The SVM (cont.).The SVM is not estimating a probability.



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)

♣♣ 3.Tuning the estimates

The smoothing parameter λ must be chosen. If the Gaussian kernel, $K(s,t) = \exp -\frac{\|s-t\|^2}{\sigma^2}$ is used then σ^2 must also be chosen. λ and σ^2 can be jointly chosen by GACV or by 5-fold crossvalidation-(next slide).

♣♣ 3. Tuning the estimates (cont.).

GACV (generalized approximate cross validation). For penalized likelihood: [XiangWahba96][XLin98];
For the (two class) SVM[Wahba99]
For the MSVM[Lee02][LeeLinWahba02].

Leaving out one for the two class SVM :

$$V_O(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[1 - y_i f_\lambda^{[i]}(t_i)]_+$$

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

$$GACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[1 - y_i f(t_i)]_+ + D(y, f_\lambda)$$

where

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

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.

♣♣ 4. The Multicategory Support Vector Machine (MSVM).

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.

♣♣ 4. The Multicategory Support Vector Machine (MSVM)(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

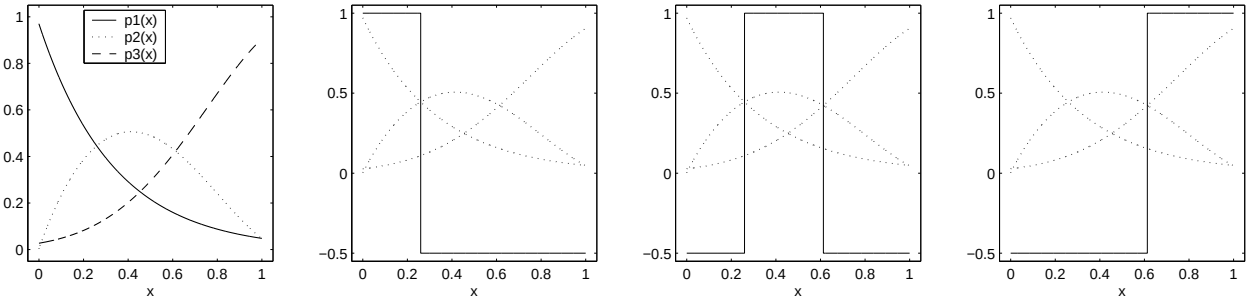
$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r\neq cat(i)} (f^r(t_i) + \frac{1}{k-1})_+ + \lambda \sum_{j=1}^{k} \|h^j\|^2_{\mathcal{H}_K}$$

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

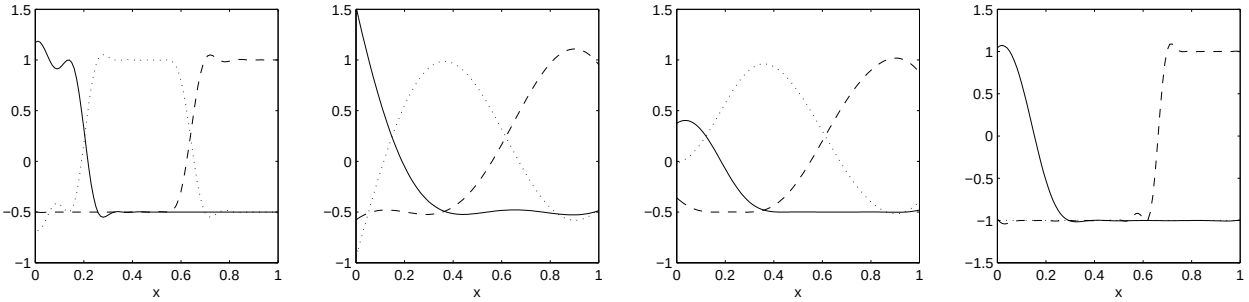
The $k = 2$ case reduces to the usual 2-category SVM.

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.

♣♣ 4. The Multicategory Support Vector Machine (MSVM)(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.

♣♣ 4. The Multicategory Support Vector Machines(MSVM)(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

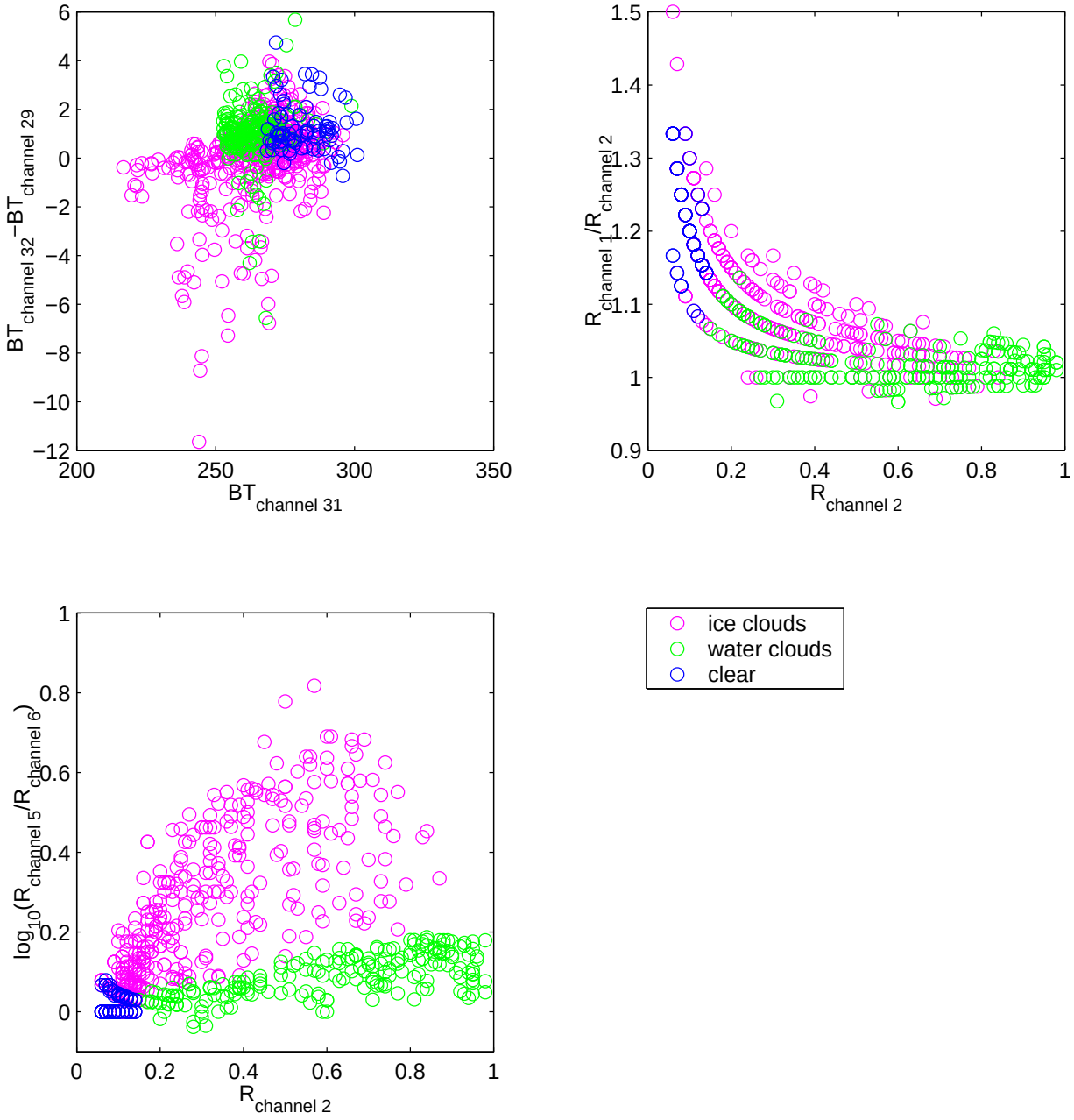
$$\sum_{\ell=1}^{k} C_{\ell j} p_\ell(x)$$

♣♣ 5. Application to the classification of upwelling MODIS radiance data to clear sky, water clouds or ice clouds.

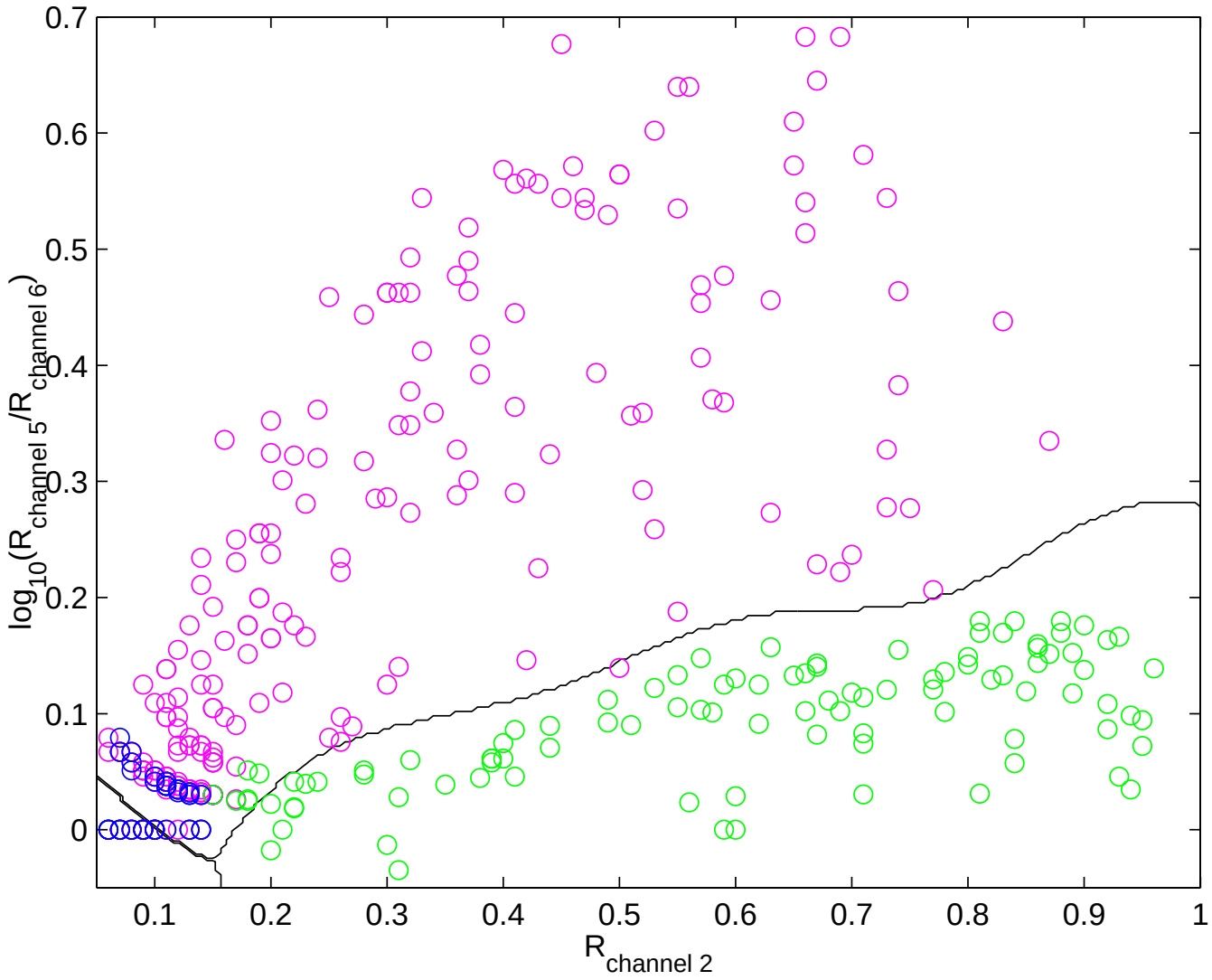
From [LWA03]. Classification of 12 channels of upwelling radiance data from the satellite- borne MODIS instrument. MODIS is a key part of the Earth Observing System (EOS).

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

744 simulated radiance profiles (81 clear-blue, 202 water clouds-green, 461 ice clouds-purple).



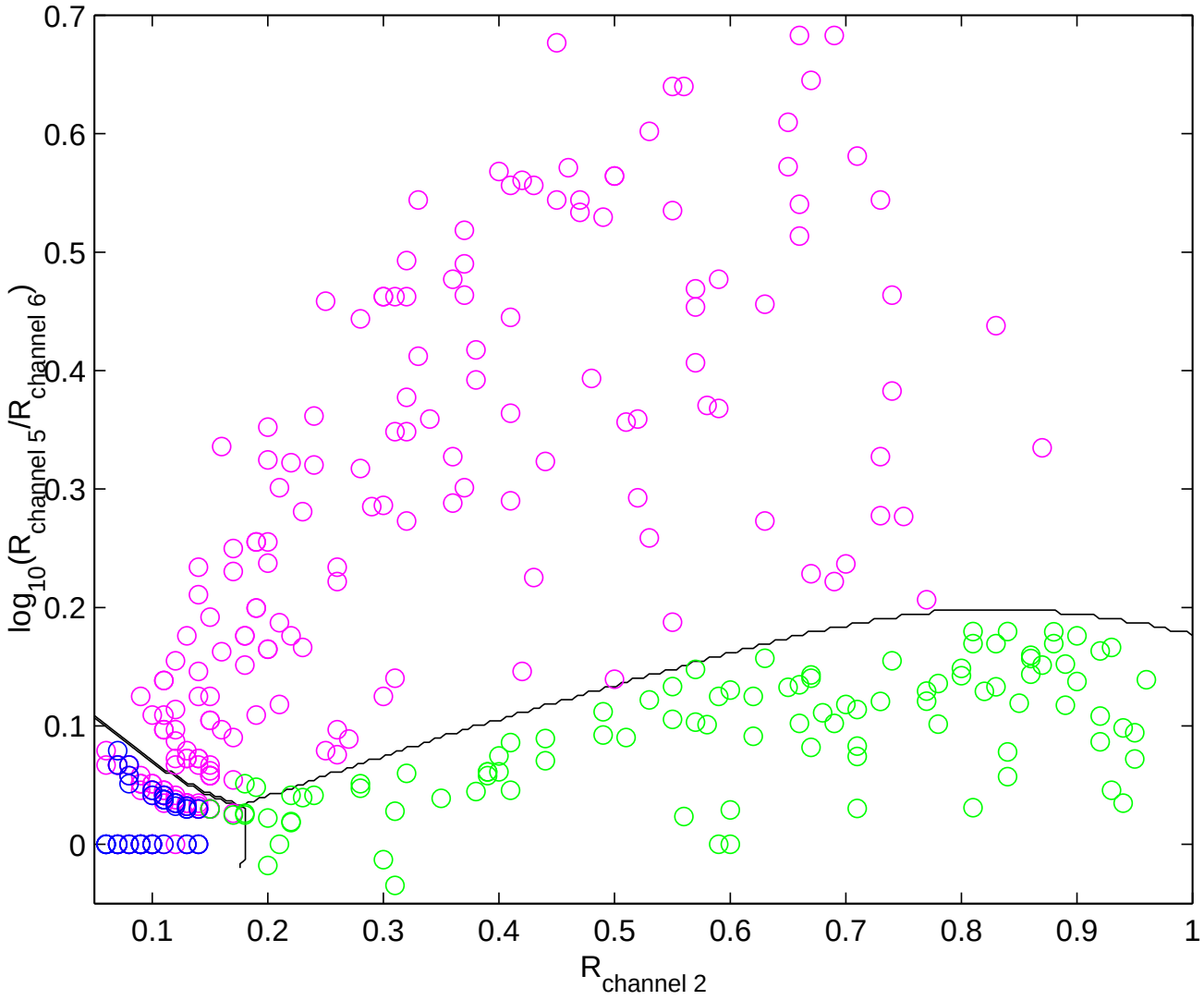
Pairwise plots of three different variables (including composite variables.(purple = ice clouds, green = water clouds, blue = clear)



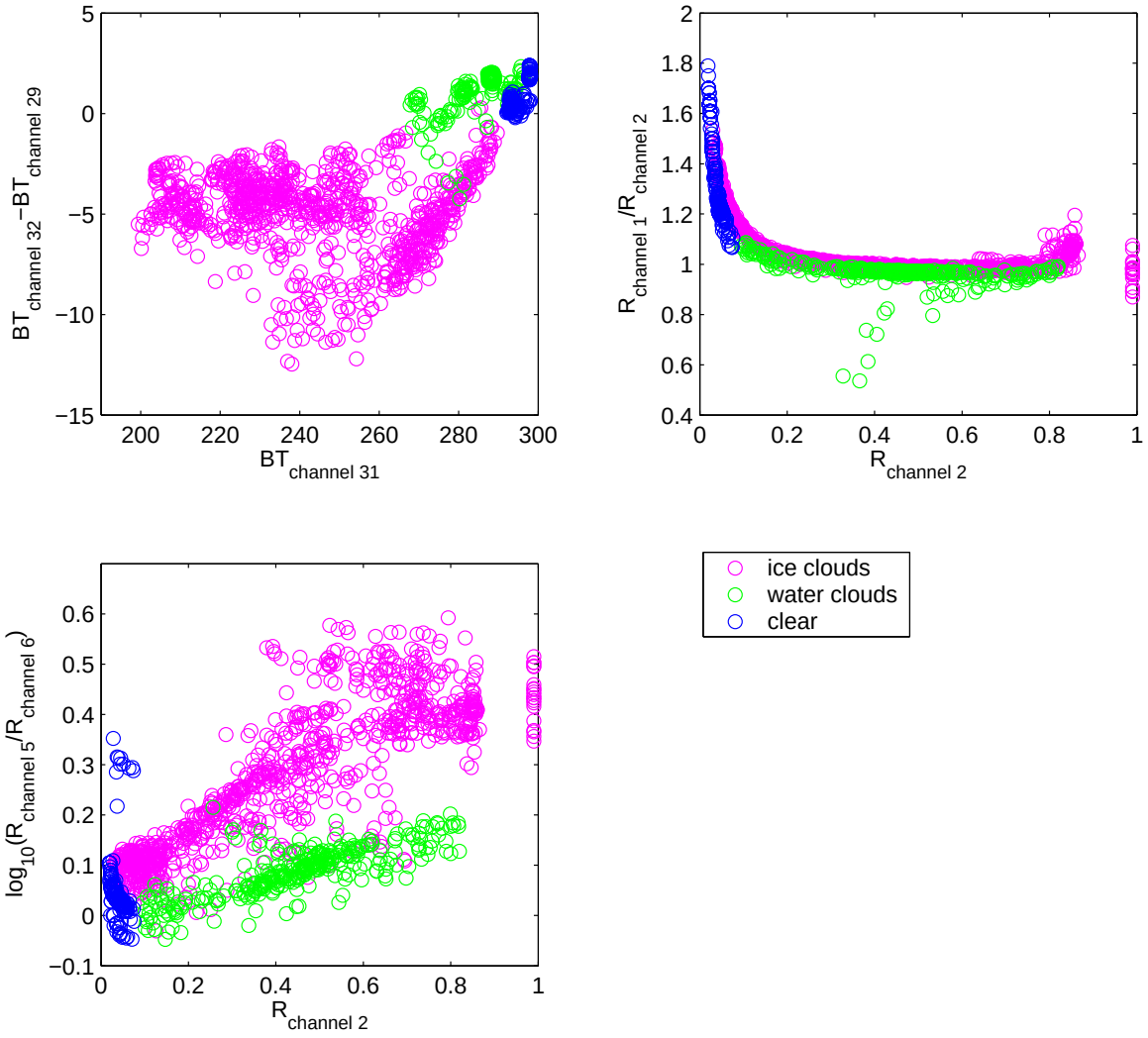
Classification boundaries on the 374 test set determined by the MSVM using 370 training examples, two variables, one is composite. Y. K. Lee Student poster prize AMetSoc Satellite Meteorology and Oceanography session.

MSVM test error rates for the combinations of variables and classifiers.

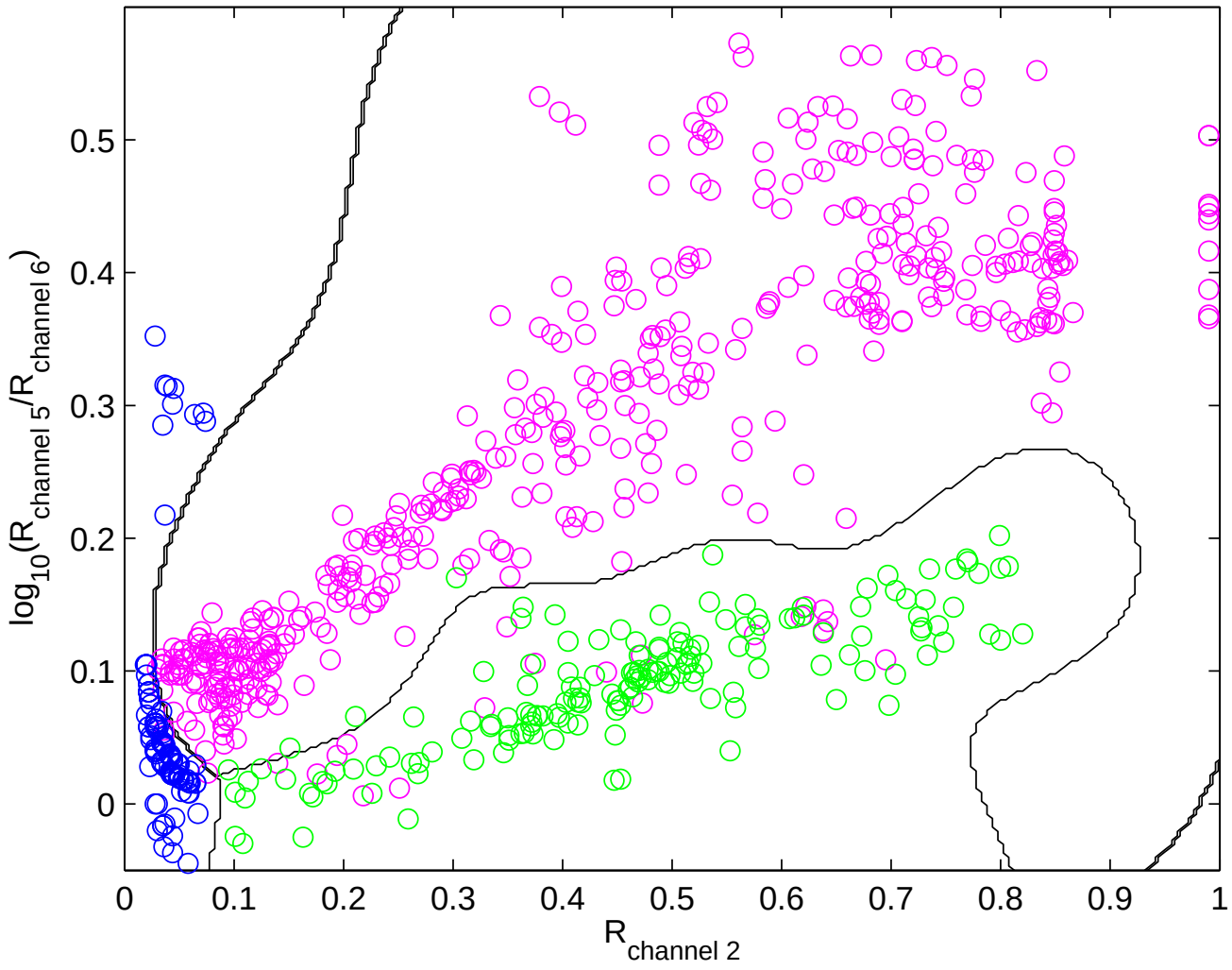
Number of variables	Variable descriptions	Err rates (%)
2	(i) R_2, $\log_{10}(R_5/R_6)$	11.50
5	(i)+R_1/R_2, BT_{31}, $BT_{32} - BT_{29}$	10.16
12	(ii) original 12 variables	12.03
12	log transformed (ii)	9.89



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



Real Data: Pairwise plots of three different variables (including composite variables). (purple = ice clouds, green = water clouds, blue = clear) 1536 profiles ”Labeled by an expert.” Note remarkable similarity to simulated data!



Real Data: Classification boundaries on the test set determined by the MSVM using training examples, two variables, one is composite.

MSVM test error rates for the combinations of variables and classifiers.

Simulated Data:

Number of variables	Variable descriptions	Err rates (%)
2	(i) R_2, $\log_{10}(R_5/R_6)$	11.50
5	(i)+R_1/R_2, BT_{31}, $BT_{32} - BT_{29}$	10.16
12	(ii) original 12 variables	12.03
12	log transformed (ii)	9.89

Real data:

Number of variables	Variable descriptions	Err rates (%)
2	(i) R_2, $\log_{10}(R_5/R_6)$	4.69
5	(i)+R_1/R_2, BT_{31}, $BT_{32} - BT_{29}$	0.26
12	(ii) original 12 variables	0.78
12	log transformed (ii)	0.65

Test error rate of the MODIS cloud masking algorithm on the real data: 18% (!)

♣♣ 6. Closing Remarks

Experimental software for the MSVM is available on a limited basis from Yoonkyung Lee yklee@stat.ohio-state.edu until Dec 31. Public code under development.

Simulated MODIS Data for the conditions studied here is reasonably realistic, and may provide a useful rough cut when real labeled training data is not available.

The tuned MSVM is amazingly good at ’learning’ how an expert labels MODIS radiance profiles.

The MSVM may be adjusted to reflect different costs for different kinds of misclassifications.

Interesting questions arise with regard to choosing important variables or combinations of variables.

The MSVM is appropriate for many classification problems.