Statistics 860 Lecture 15

Multicategory Support Vector Machines

References

Y. Lee and C. K. Lee Classification of Multiple Cancer Types by Multicategory Support Vector Machines Using Gene Expression Data. Bioinformatics 19 (2003) 1132-1139. `lee.lee.pdf`

Y. Lee, Y. Lin and G. Wahba. Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data, JASA, March 04
`lee.lin.wahba.04.pdf`

Y. Lee, G. Wahba and S. Ackerman. Classification of Satellite Radiance Data by Multicategory Support Vector Machines, JTECH Feb 04
`lee.wahba.ackerman.04.pdf`
`lee.wahba.ackerman.corr.04.pdf`.

X. Lin. *Smoothing Spline Analysis of Variance for Polychotomous Response Data*. PhD thesis, Department of Statistics, Uiversity of Wisconsin, Madison WI, 1998i, TR 1003. `xiwuth.pdf`.

G. Wahba. Soft and hard classification by reproducing kernel Hilbert space methods. Proc. NAS 2002, 99, 16524-16503. `tr1067.pdf`

♣♣ Multichotomous penalized likelihood[xiwuth.pdf].

$k+1$ 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 [xiwuth.pdf]: 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.

(Also J. Zhu and T. Hastie, Biostatistics 5:427-443, 2003- $f_j(t) = (b_j, x)$ with $\sum_{j=1}^{K} f_j(t) = 0$ constraint)

♣♣ 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

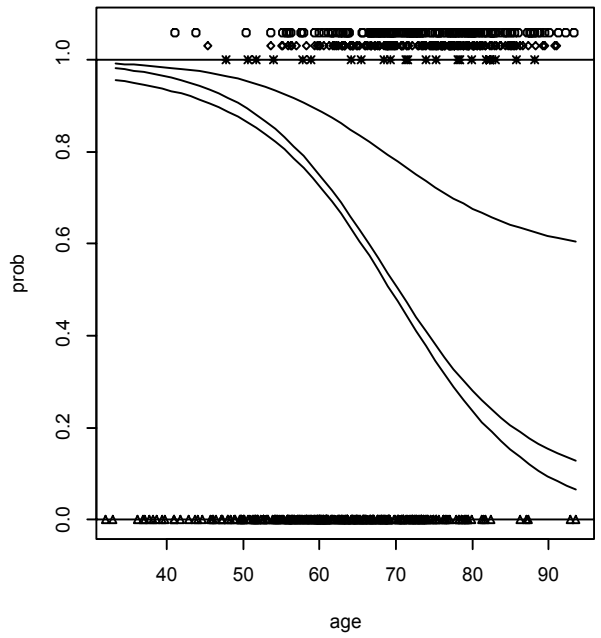
$$\sum_{j=1}^{k} \lambda_j \|h^j\|^2_{\mathcal{H}_K},$$

and the optimization problem becomes: Minimize

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

♣♣ 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[`xiwuth.pdf`].



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.)

♣♣ Multicategory support vector machines (MSVMs).

From [lee.lin.wahba.04.pdf], [lee.wahba.ackerman.corr.04.pdf], 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.

♣♣ 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

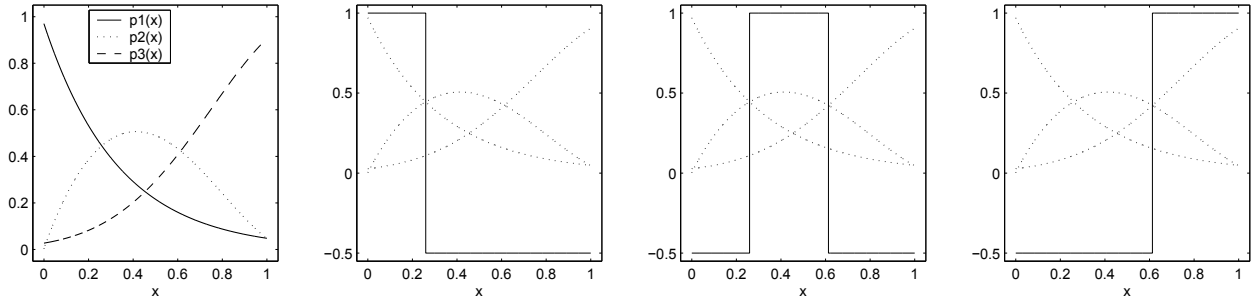
$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r\neq cat(i)} (f^r(t_i) + \frac{1}{k-1})_+ + \lambda \sum_{j=1}^{k} \|h^j\|^2_{\mathcal{H}_K}$$

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

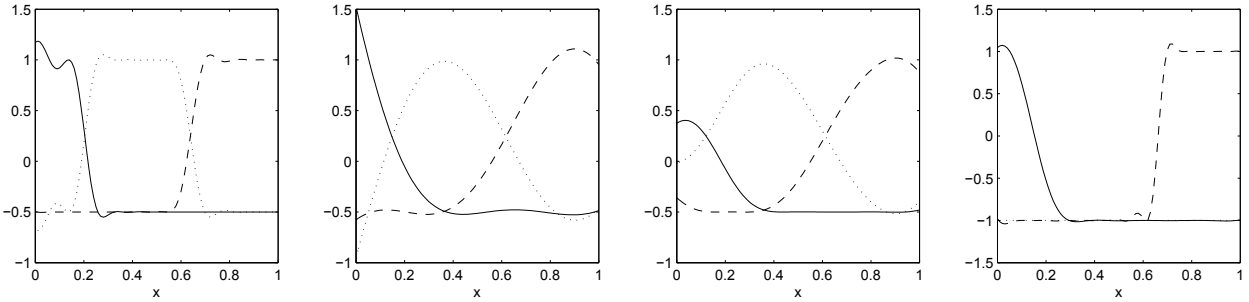
The $k = 2$ case reduces to the usual 2-category SVM.

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.

♣♣ 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.

♣♣ 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

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

♣♣ Tuning the estimates.

GACV (generalized approximate cross validation).
Penalized likelihood:
[xiang.wahba.sinica.pdf][lin.xiwuth.ps];
SVM[nips97rr.ps, nips97rr.typos.ps],
MSVM[lee.lee.pdf][lee.lin.wahba.04.pdf].

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

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

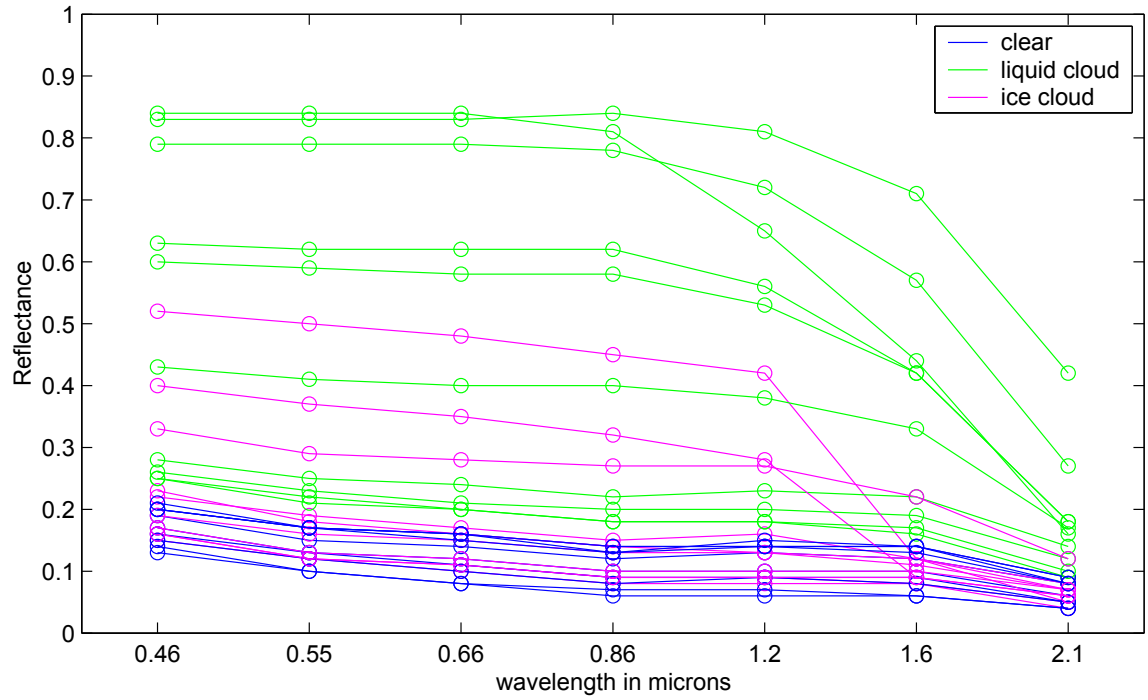
is obtained by a tailored perturbation argument. Easy to compute for the SVM, use randomized trace techniques to estimate the perturbation in the likelihood case.

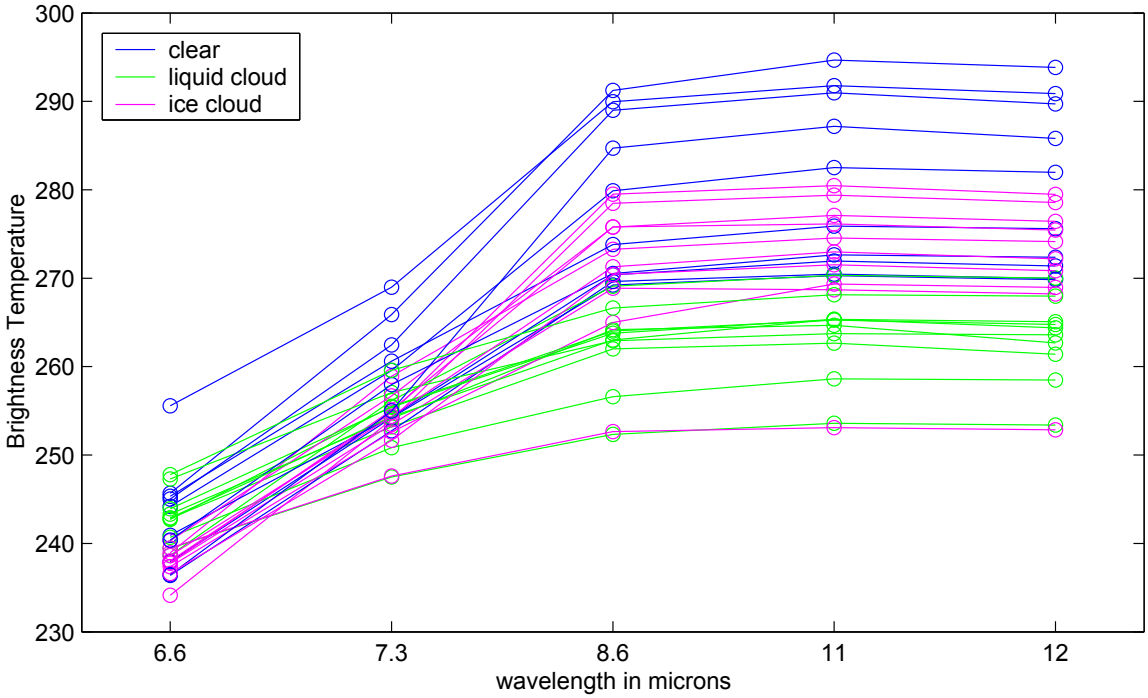
♣♣ 8. The classification of upwelling MODIS radiance data to clear sky, water clouds or ice clouds.

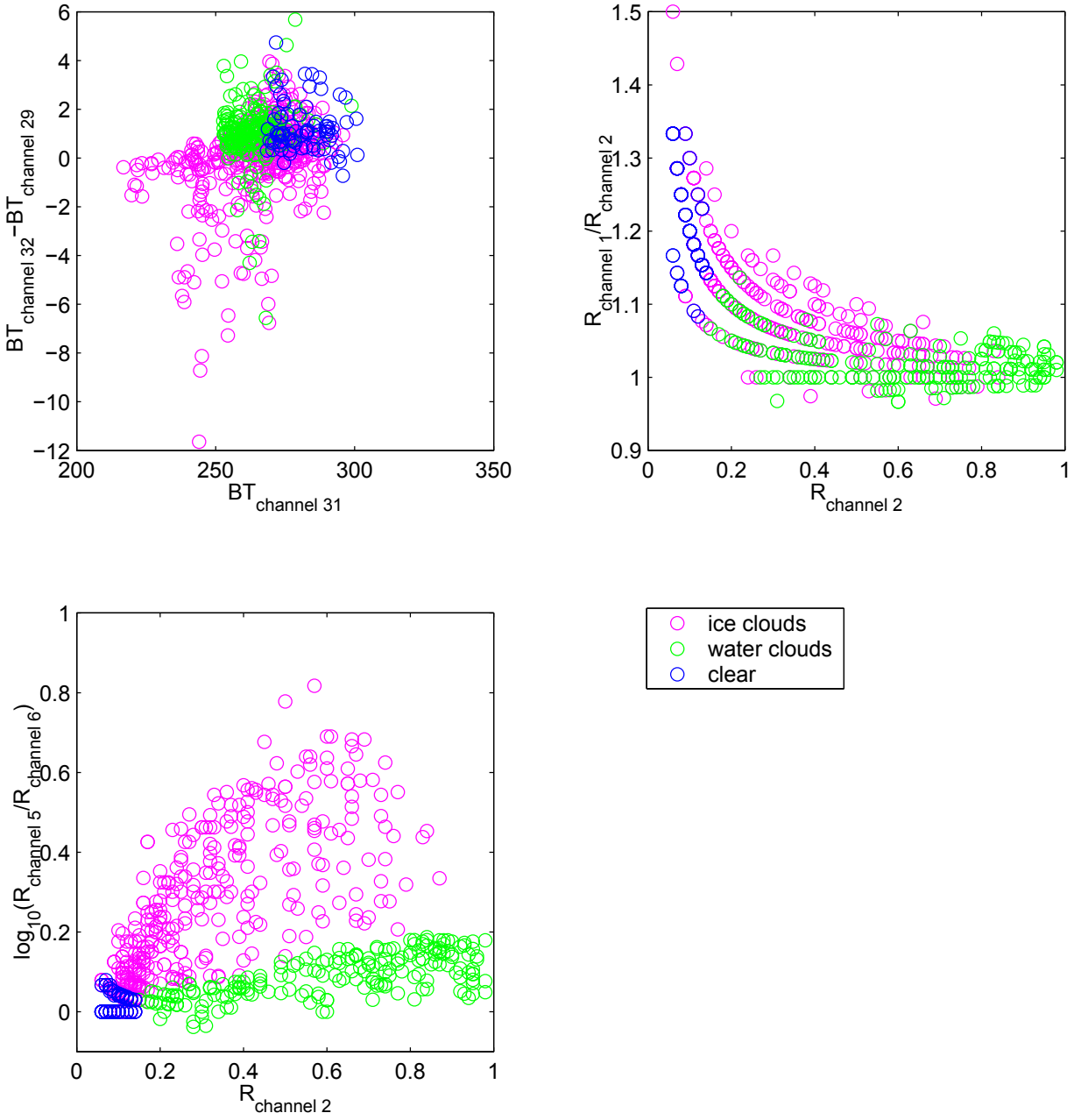
From [lee.wahba.ackerman.04.pdf].Classification of 12 channels of upwelling radiance data from the satellite- borne MODIS instrument. MODIS is a key part of the Earth Observing System (EOS).

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

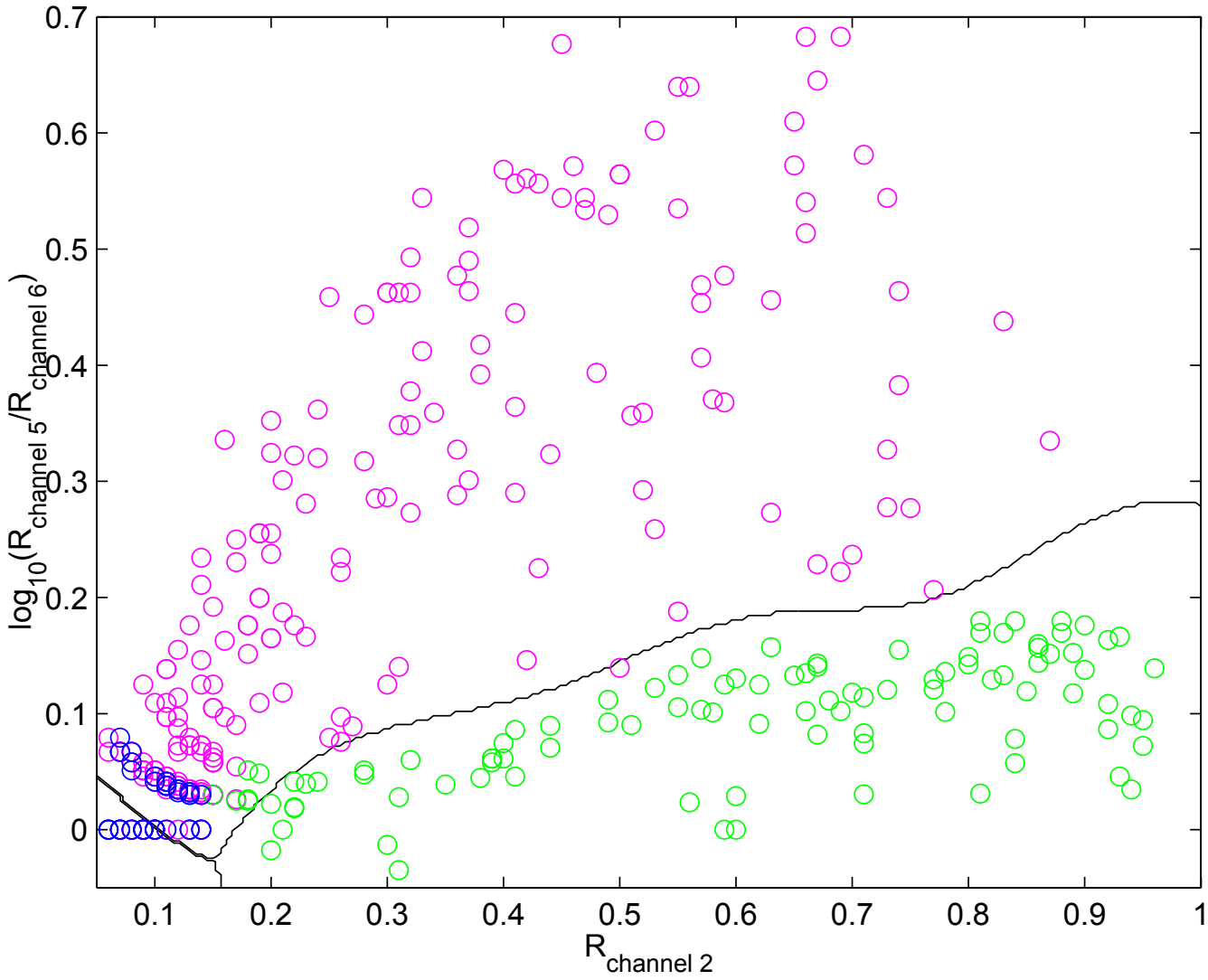
Next page: 744 simulated radiance profiles (81 clear-blue, 202 water clouds-green, 461 ice clouds-purple). 10 samples from clear, from water and from ice:


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

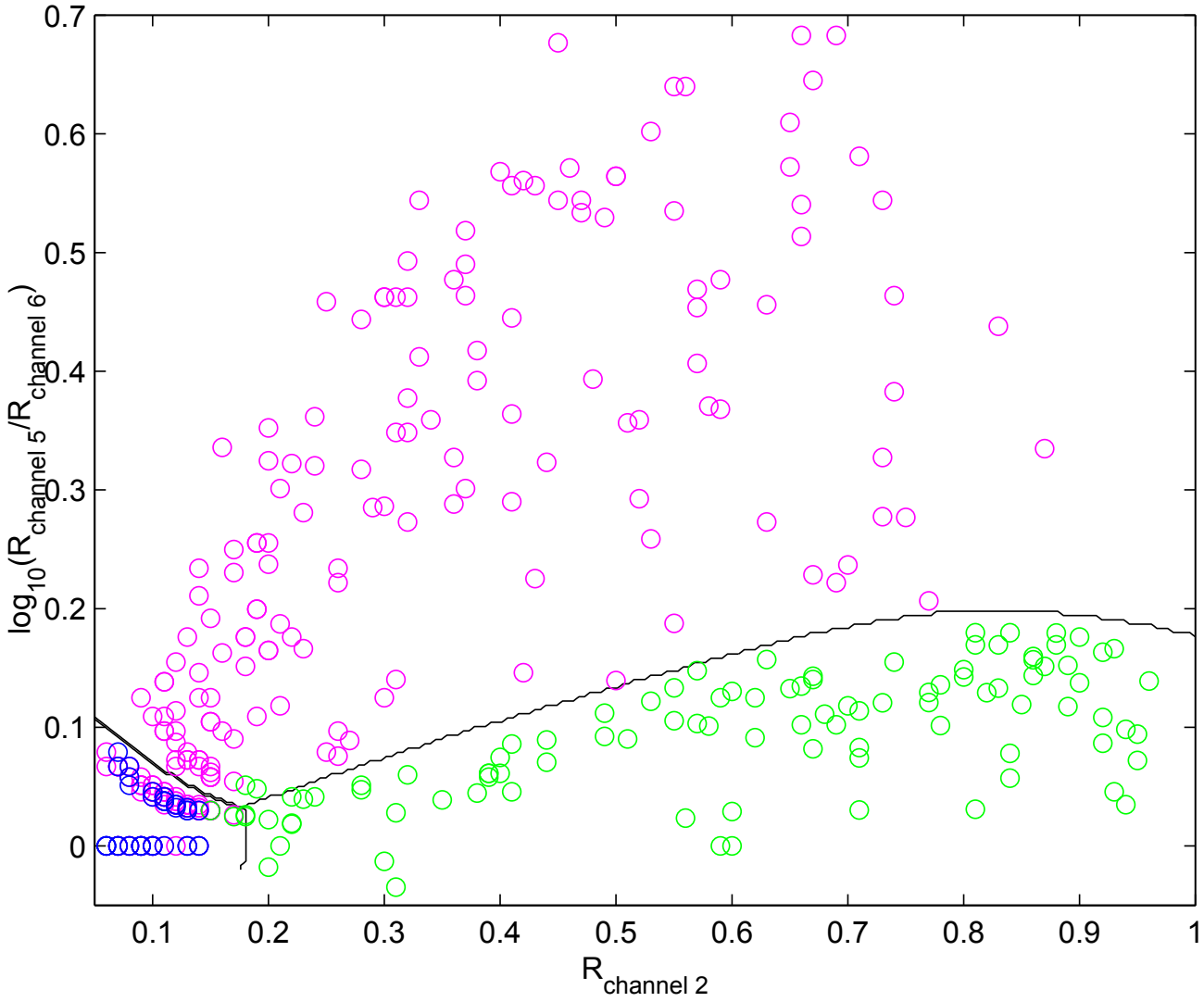

clear
liquid cloud
ice cloud
Brightness Temperature
wavelength in microns
230
240
250
260
270
280
290
300
6.6
7.3
8.6
11
12



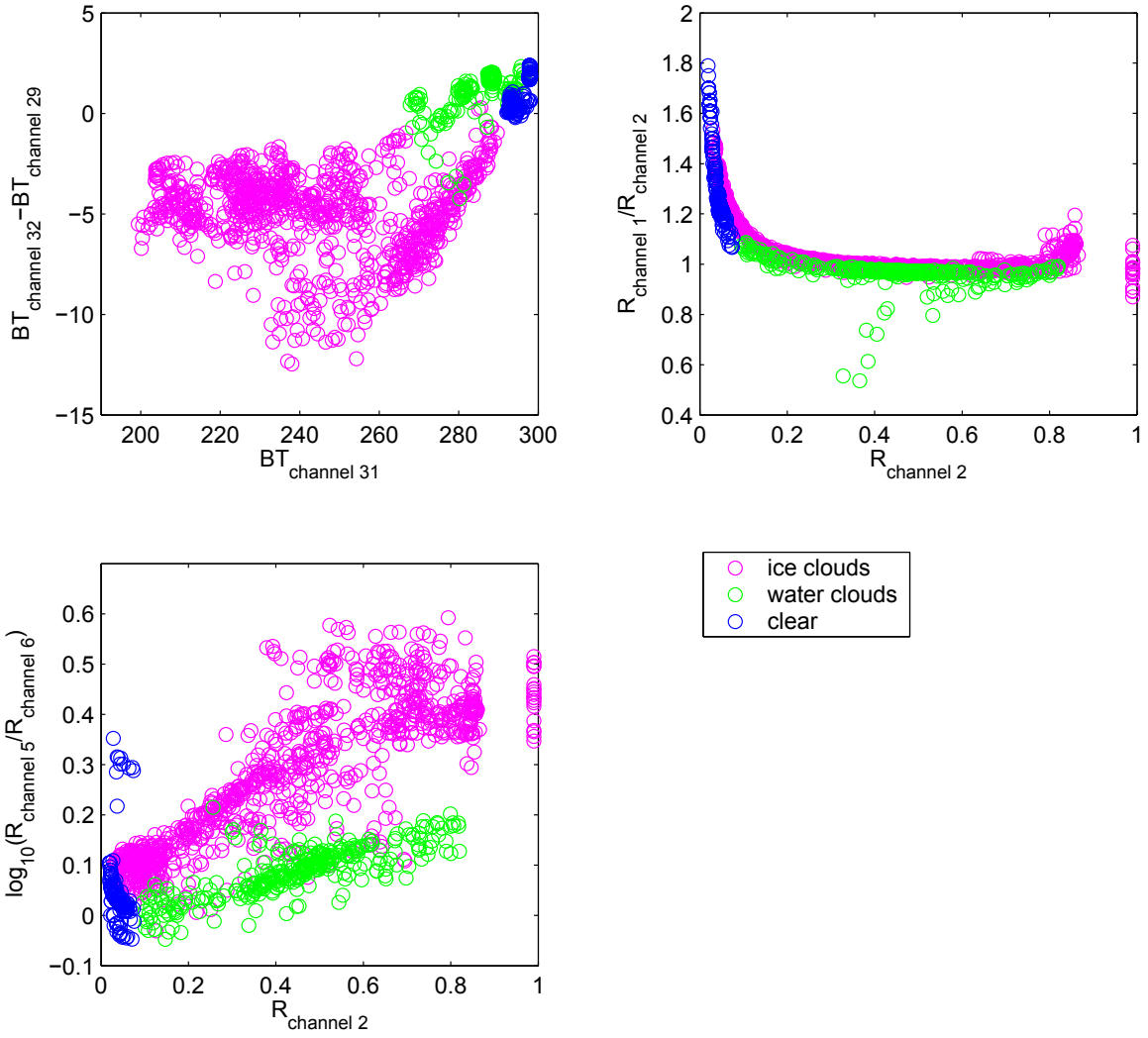
Pairwise plots of three different variables (including composite variables.(purple = ice clouds, green = water clouds, blue = clear)



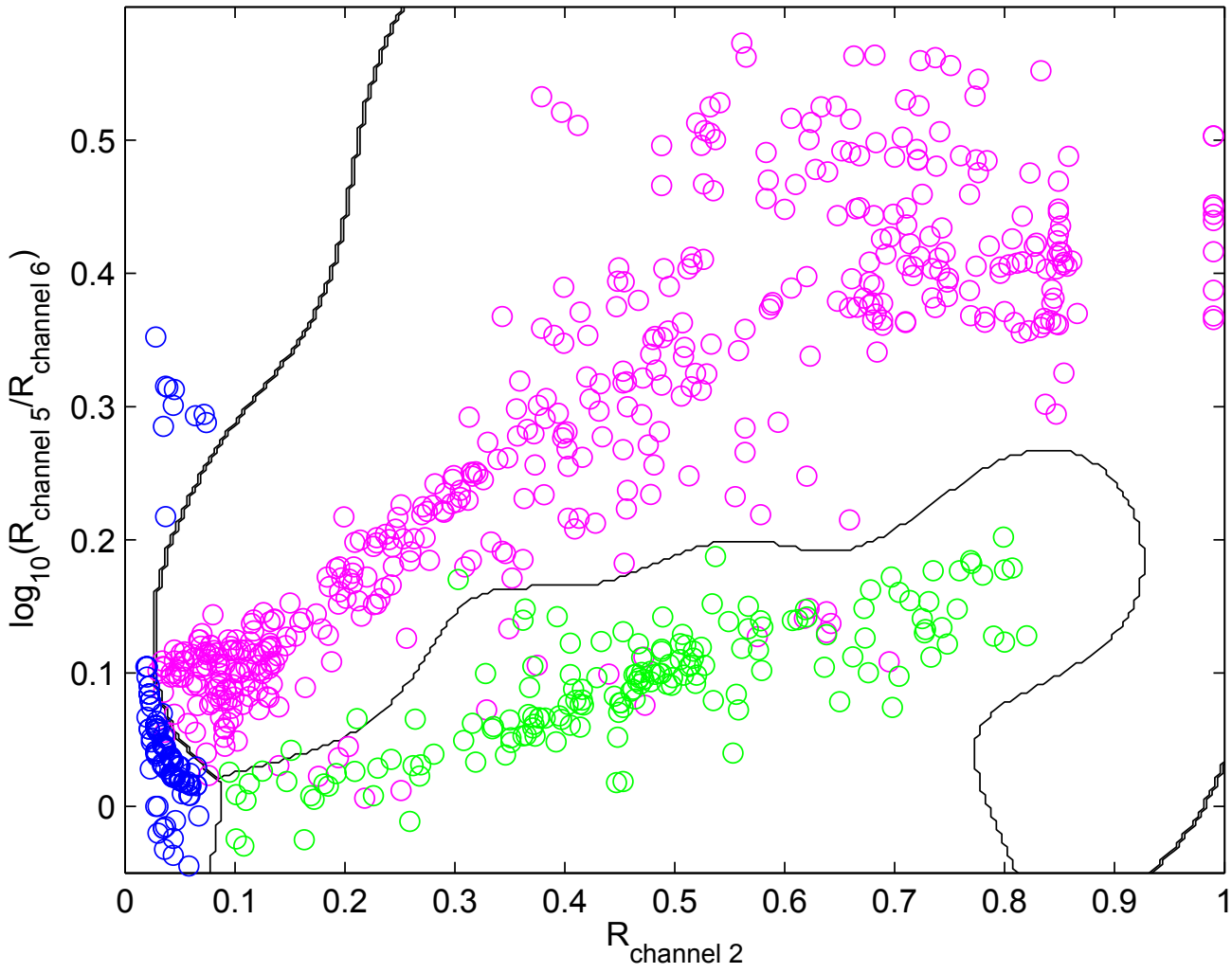
Classification boundaries on the 374 test set determined by the MSVM using 370 training examples, two variables, one is composite. Y. K. Lee Student poster prize AMetSoc Satellite Meteorology and Oceanography session.



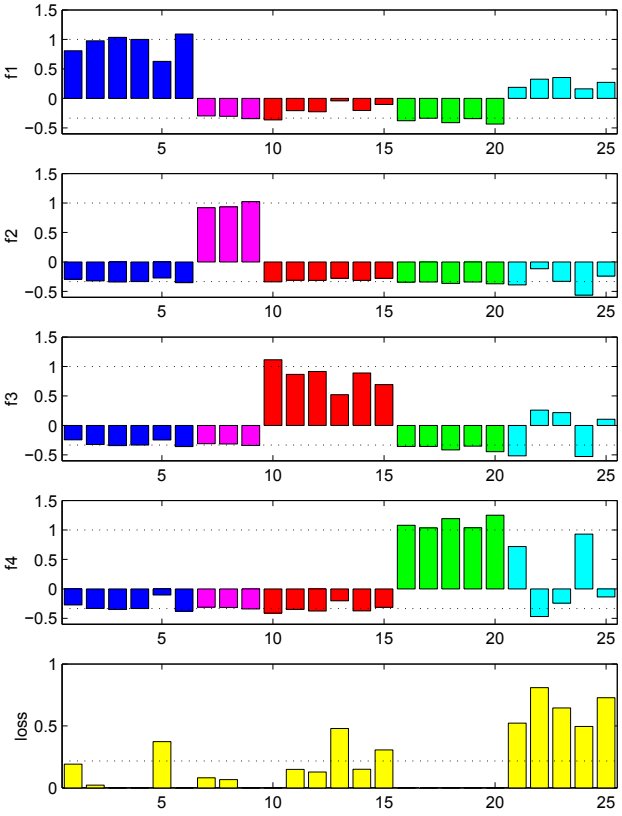
Classification boundaries determined by the nonstandard MSVM when the cost of misclassifying clear clouds is 4 times higher than other types of misclassifications.



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.



The first four panels show the predicted decision vectors (f_1, f_2, f_3, f_4) at the test samples. The four class labels are coded according as EWS in blue: $(1, -1/3, -1/3, -1/3)$,
BL in purple: $(-1/3, 1, -1/3, -1/3)$,
NB in red: $(-1/3, -1/3, 1, -1/3)$,

and RMS in green: $(-1/3, -1/3, -1/3, 1)$. The colors indicate the true class identities of the test samples. We can see from the plot that all the 20 test examples from 4 classes are classified correctly and the estimated decision vectors are pretty close to their ideal class representation. The fitted MSVM decision vectors for the 5 non SRBCT samples are plotted in cyan. The last panel depicts the loss for the predicted decision vector at each test sample. The last 5 losses corresponding to the predictions of non SRBCTs all exceed the threshold (the dotted line) below which means a strong prediction. Three test samples falling into the known four classes can not be classified confidently by the same threshold.

How strong is the classification?

A decision vector close to a class code in the multi-class case may mean a strong prediction. Recall the multiclass hinge loss for an observation y_i at x_i:

$$\mathcal{C}(y_i, f(x_i)) = \sum_{r=1}^{k} L_{cat(y_i)r}(f^r(x_i) - y_{ir})_+$$

measures the proximity between an MSVM decision vector and a coded class.

Cross validation heuristics will be used to estimate strength of a prediction, (standard case). For each i, leaving out the ith example, collect

$$\mathcal{C}(\widehat{y}_i, f^{[-i]}(x_i)) \equiv \mathcal{C}^{[-i]}$$

where $\widehat{y}_i$ is the coded version of the identification made by $f^{[-i]}$, along with an indicator as to whether $\widehat{y}_i = y_i$, i. e. whether the correct identification was made by $f^{[-i]}$.

Heuristically, with some symmetry assumptions, all of the $\mathcal{C}^{[-i]}$ associated with a correct classification are then pooled, and all of the $\mathcal{C}^{[-i]}$ associated with an incorrect classification are pooled and can be used to form an estimate of the probability of correct classification, as a function of $\mathcal{C}(y, f(x))$ for future observations. When the training set is completely correctly classified, then the 95% of the $\mathcal{C}^{[-i]}$ distribution could be used.