Mathematical Programming in Machine Learning and Data Mining
January 14-19, 2007
Banff International Research Station

Grace Wahba

On ‘Consuming’ Mathematical Programming: Selection of High Order Patterns in Demographic and Genetic Data

All papers/preprints up at
`http://www.stat.wisc.edu/~wahba`
→ TRLIST

- “LASSO-Patternsearch Algorithm with Application to Ophthalmalogy Data”, Weiliang Shi, Grace Wahba, Stephen J. Wright, Kristine Lee, Ronald Klein and Barbara Klein, TR 1131, October 2006.

1.01.07

The LASSO-Patternsearch Algorithm involves a smooth nonlinear optimization problem with a potentially very large number of variables with bound constraints but many of which will be 0. Steve Wright devised the code for this problem in conjunction with Weiliang Shi, and we have applied it to examine clusters or patterns of risk factors for an eye disease to data from the Beaver Dam Eye Study (BDES), with interesting scientific results. I'll describe the statistical setting of the algorithm and the scientific results from BDES. We also applied it to the analysis of SNP data and covariates in a case control study with excellent results. This problem potentially has a humongous number of variables, but we pruned it to fit into the code, which presently can deal with 4000 unknowns on our 2G system.

I’ll briefly mention two other papers whose core is a convex cone algorithm:

- “A Framework for Kernel Regularization With Application to Protein Clustering”, by Fan Lu, Sündüz Keleş, Stephen J. Wright and Grace Wahba, PNAS, 102(August 2005).

- “Robust Manifold Unfolding with Kernel Regularization” by Fan Lu, Yi Lin and G. Wahba, TR 1108, October 2005.

Background for the LASSO-Patternsearch: Review of parametric logistic regression:

Data $\{y_i, x(i)\}, i = 1, \cdots, n$

y_i - response of ith subject with p attributes:

$$x(i) = (x_1(i), x_2(i), \cdots, x_p(i))$$

$y_i, x_1(i), x_2(i), \cdots, x_p(i) \in \{0, 1\}$.

Let

$$p(x) = Prob(y = 1|x)$$

and

$f(x) = log(p(x)/(1-p(x))) \quad [the\; log\; odds\; ratio]$

then

$$p(x) = e^{f(x)}/(1 + e^{f(x)}).$$

Parametric logistic regression, continued.

The logit $f(x)$ is modeled as

$$f(x) = \sum_{\ell=0}^{N} c_\ell B_\ell(x)$$

where the B_ℓ are given basis functions. The B_ℓ depend on $x = x_1, ..., x_p$ in some specified way. The $\{c_\ell\}$ are found by minimizing the negative log likelihood:

$$\mathcal{L}(y, f) = \sum_{i=1}^{n} -y_i f(x(i)) + log(1 + e^{f(x(i))}).$$

MATLAB, SAS, etc. return estimates of the coefficients $\{c_\ell\}$ along with confidence intervals and p-values. Want N no more than around 10 or 20.

In the special case $B_0(x) = 1, B_\ell(x) = x_\ell, \ell = 1, ..., p$, then $N = p + 1$ and

$$f(x) = c_0 + \sum_{\ell=1}^{p} c_\ell x_\ell.$$

Whatever the B_ℓ, then c_ℓ is the log odds ratio conditional on $B_m(x) = 0$ for m not equal to ℓ: :

$$e^{c_\ell} = OR \;\; = \;\; \frac{Prob\,(y = 1|B_\ell(x) = 1, B^{-\ell}(x) = 0)}{Prob\,(y = 0|B_\ell(x) = 1, B^{-\ell}(x) = 0)} / \frac{Prob\,(y = 1|B_\ell(x) = 0, B^{\ell}(x) = 0)}{Prob\,(y = 0|B_\ell(x) = 0, B^{\ell}(x) = 0)}$$

where

$$B^{-\ell}(x) \;\; = \;\; \{B_m(x)\}, m \neq \ell.$$

LASSO-Patternsearch procedure involves a large to very large number N of basis functions, possibly $N \geq n$, by minimizing the ℓ_1 *penalized likelihood*

$$\mathcal{L}(y, f) + \lambda \sum_{\ell=1}^{N} |c_\ell|.$$

with

$$f(x) = \sum_{\ell=0}^{N} c_\ell B_\ell(x).$$

The so-called ℓ_1 penalty $\sum_{\ell=1}^{N} |c_\ell|$ has the property that many smaller c_ℓ will be set to 0, depending on λ. For given λ, the c_ℓ can be found numerically in MATLAB for moderate size N, but that is not good enough.

For the LASSO-Patternsearch the basis functions will be all products of the x_r up to order q:

$$B_{j_1,j_2,..,j_r}(x) = \prod x_{j_1}x_{j_2}...x_{j_r}, r = 1, \cdots, q.$$

Thus, $B_{j_1,j_2,...,j_r}(x) = 1$ if x is a p-vector which has ones in each of the $j_1, j_2, \cdots, j_r$ positions, and $B_{j_1,...,j_r}(x) = 0$ otherwise. The number N of basis functions is then

$$N = \binom{p}{0} + \binom{p}{1} + \binom{p}{2} + ... + \binom{p}{q}.$$

For $q = p$, (all possible patterns), $N = 2^p$.

Note that the conditional distribution of one Bernoulli random variable y given p other Bernoulli random variables $x_1, \cdots, x_p$ has 2^p paramteters and can be expanded in complete generality in these basis functions. The representation will be most compact, however, if all the risky variables are coded with the risky direction as 1.

Now

$$\begin{aligned} f(x) = c_0 + \sum c_\alpha B_\alpha(x) + \sum c_{\alpha\beta} B_{\alpha,\beta}(x) \\ + \cdots \sum c_{j_1,j_2,\cdots,j_q} B_{j_1,j_2,\cdots,j_q}(x) \\ \equiv c_0 + \sum_{\ell=1}^{N} c_\ell B_\ell(x). \end{aligned} \tag{1}$$

The LASSO-Patternsearch has the following steps:

Step 1. Minimize $\mathcal{L}(y, f) + \lambda \sum_{\ell=1}^{N} |c_\ell|$, choose λ by GACV.

Step 2. Enter all basis functions with $\ell : |c_\ell| > 0$ into a parametric logistic regression model:

$$f(x) = \sum_{\ell: c_\ell > 0} a_\ell B_\ell(x)$$

and fit.

Step 3. Select all ℓ for which a_ℓ are significant at the q% level, to fit the final model:

$$f(x) = \sum_{\ell: a_\ell \ \ significant} b_\ell B_\ell(x).$$

q is another tuning parmeter, chose q by BGACV. Examine the resulting patterns (B_ℓ)s with significant b_ℓ's.

Step 4. Interpret, demonstrate ”significance after datamining”.

Application to progression of myopia from the Beaver Dam Eye Study, BDES 1 to BDES2, $n = 876$ records of persons aged 60-69 at BDES1. A person whose 'worse eye' scored at a decrease of .75 Diopters or more is labeled $y = 1$, and 0 otherwise. About 13% of this group was scored $y = 1$.

Table 1: Trial Variables and Cutpoints

	variable	description	binary cut point (higher risk $X = 1$)
X_1	sex	sex	Male
X_2	inc	income	$<$ 30
X_3	jomyop	juvenile myopia	$<$ 21
X_4	catct	cataract	4-5
X_5	iop	intraocular pressure	22+
X_6	pky	packyear	$>$30
X_7	asa	aspirin	not taking
X_8	vtm	vitamin	not taking

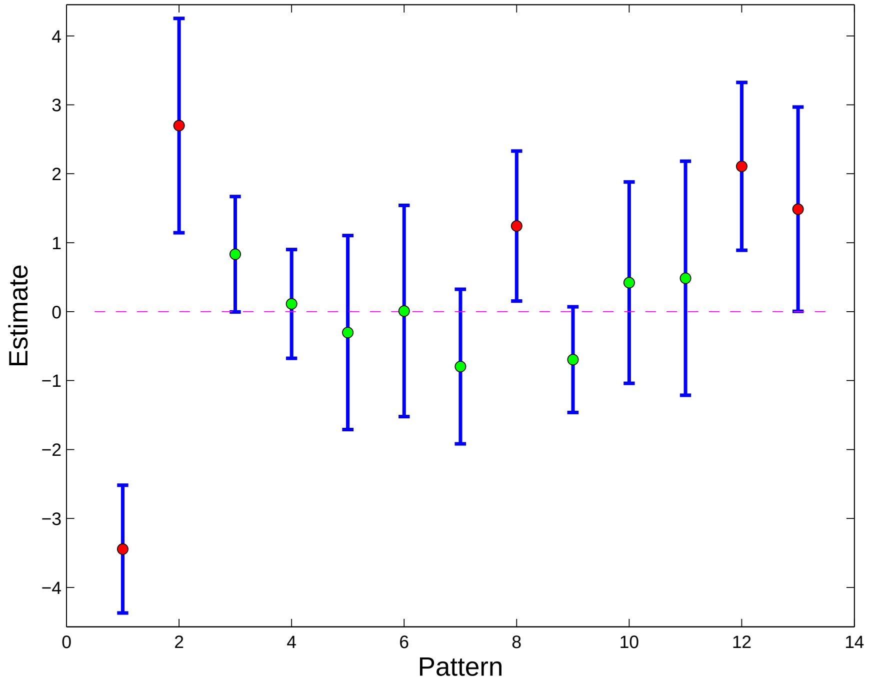
Step 1. $p = 7$ variables, $q = p = 7$, $N = 2^7$ basis functions, minimize $\mathcal{L}(y, f) + \lambda \sum_{\ell=1}^{127} |c_\ell|$, choose λ by GACV.

Step 2. Parametric logistic regression for patterns surviving Step 1:

Table 2: Patterns Surviving Step 1

pattern				odds ratio
1. constant				0.032
2. catct				14.861
3. asa				2.302
4. vtm				1.119
5. inc	pky			0.738
6. catct	asa			1.009
7. catct	vtm			0.451
8. pky	vtm			3.462
9. sex	asa	vtm		0.498
10. inc	catct	pky		1.522
11. inc	pky	vtm		1.624
12. sex	inc	jomyop	asa	8.224
13. sex	inc	catct	asa	4.419

Step 2. Confidence intervals, patterns after Step 1:



Pattern Number

Significant patterns after Step 2:

1. Constant
2. catct (Cataract)
8. pky vtm (Packyear $>$ 30 and not taking vitamins)
12. sex inc jomyop asa (Male, low income, juvenile myopia, not taking aspirin)
13. sex inc catct asa (Male, low income, cataract, not taking aspirin)

Step 3. Select all ℓ for which a_ℓ are significant at the $q\% = 96.92\% (BGACV)$ level, to fit the final model:

$$f(x) = \sum_{\ell : a_\ell \;\; significant} b_\ell B_\ell(x).$$

The (refitted) model is

$$f(catct, pky, vtm, sex, inc, jomyop, asa)$$

$$-3.29 + 2.42 * cact + 1.18 * pky * vtm$$

$$+ 1.84 * sex * inc * jomyop * asa + 1.08 * sex * inc * cat * asa.$$

The estimated log odds ration for various subgroups can be computed, for example the worst subgroup is low income male smokers with cataracts and juvenile myopia and not taking either vitamins or aspirin:

$$f = -3.29 + 2.42 + 1.18 + 1.84 + 1.08 = 3.23,$$

$$p = e^{3.23}/(1 + e^{3.23}) = .96,$$

compared to a 13% rate in the overall population.

Step 4. Having done some ”data mining”, the investigators can go back and look at classes of people who may not have been examined separately before. For example:

catct	pky	not take vitamins	risk of progression
1	1	1	17/23 = 0.7391
1	1	0	7/14 = 0.5000
0	1	1	22/137 = 0.1606
0	1	0	2/49 = 0.0408
1	0	1	18/51 = 0.3529
1	0	0	19/36 = 0.5278
0	0	1	22/363 = 0.0606
0	0	0	13/203 = 0.0640

Looking at the smokers: smokers with cataract are relatively protected by taking vitamins, and smokers without cataract are also relatively protected by taking vitamins. For non smokers taking or not taking vitamins makes no (significant) difference.

Physiologically meaningful - recent literature suggests:

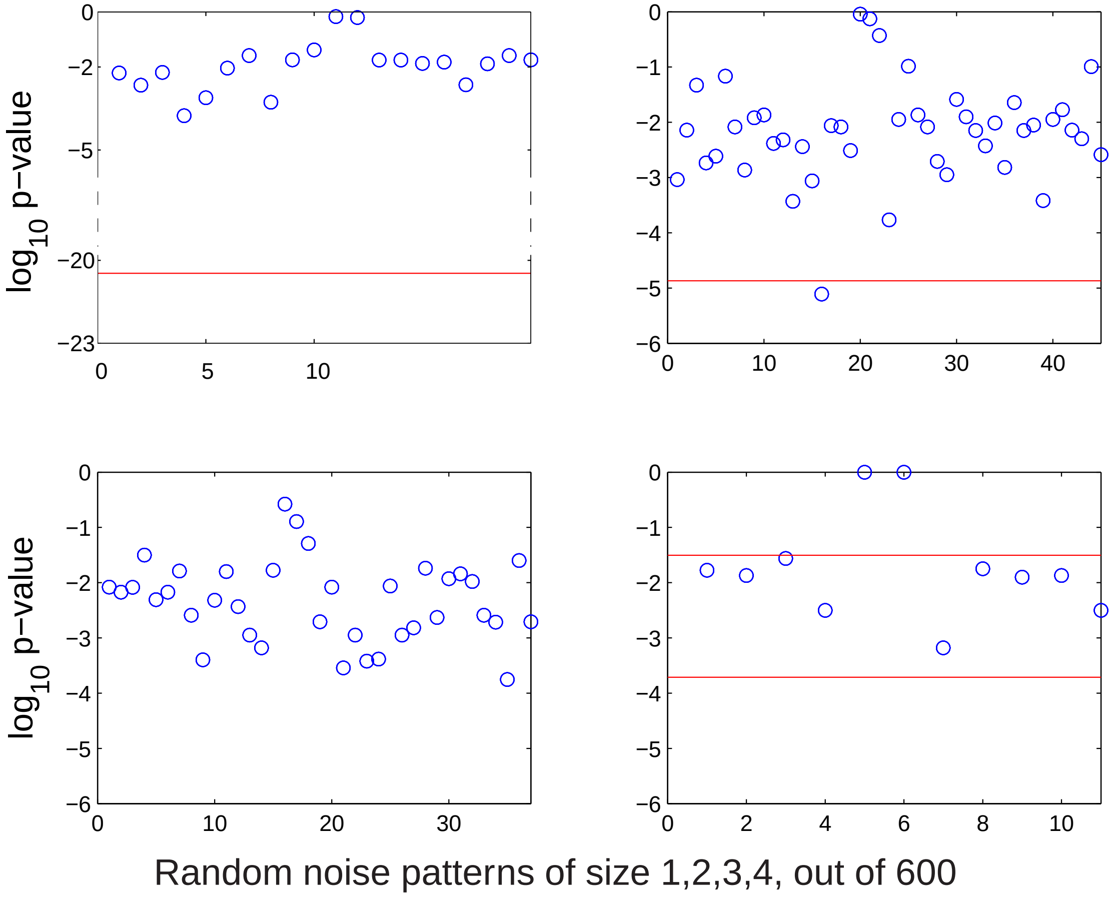
a) Certain vitamins are good for eye health.
b) Smoking depletes the serum and tissue vitamin level, especially Vitamin C and Vitamin E.

(Although as usual, a ”randomized controlled clinical trial would provide the best evidence of any effect of vitamins on progression of myopia in smokers”)

To check on the “significance” of the patterns, randomly scramble the ys while keeping the x’s fixed, and apply the entire LASSO-Patternsearch algorithm to see how often false patterns are generated. Repeat 600 times. (Statistical theory is not clear on properties of multistep procedures)

Detection of noise patterns found in scrambled data compared to observed p values:

Log p values of the patterns found (out of 600) are plotted (l. to r. then top to bottom) for observed patterns of size 1,2,3,4. Red lines are for the observed p-values for $catct$, $pky\ vtm$, none, and $sex\ inc\ jomyop\ asa$ (lower) and $sex\ inc\ catct\ asa$ (upper). Upper red line suggests that $sex\ inc\ catct\ asa$ is borderline significant.



Genetic Data (realistic simulation, not ours, not released yet)

y = phenotype, x = SNPs, alleles, covariates, after coding as 1 or 0, 9192 variables.

Train: 1500 cases, 2000 controls
Tune: 1500 cases, 2000 controls
Test: 1500 cases, 2000 controls.

Pre-screen step: 9192 variables reduced to $N = 2559$ basis functions for the LASSO step. Final model has 8 main effects and 3 interactions. Using $p = .5$ as a classifier, a competitive 12.6% error rate was obtained. Identified a SNP near most of the genes that were used to generate the data.

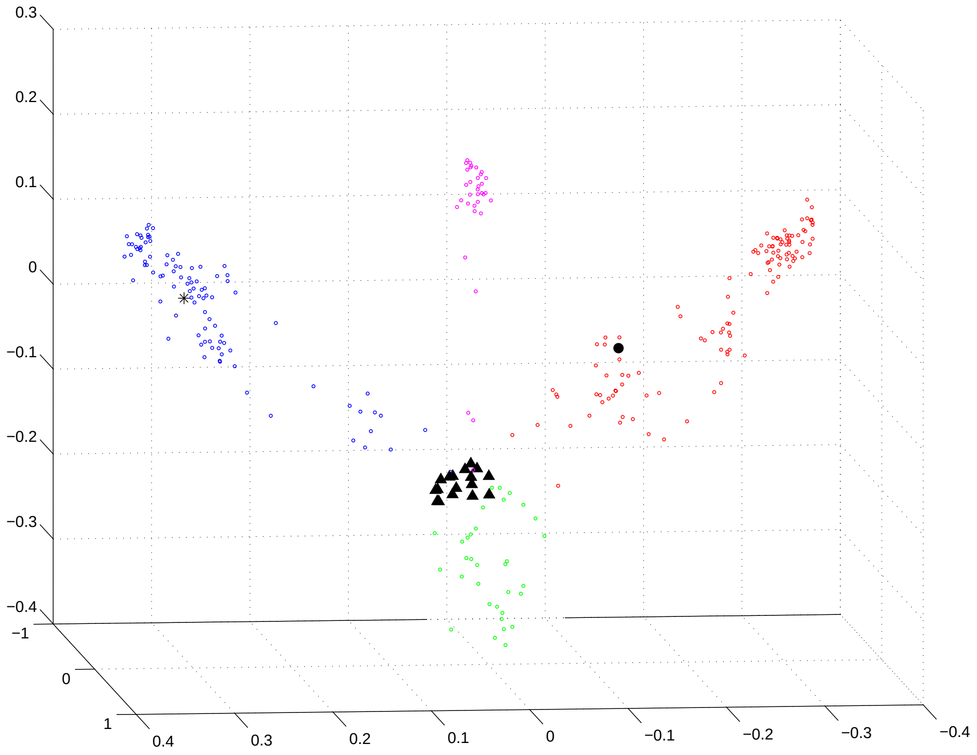
Regularized Kernel Estimation (RKE)
Lu, Keles, Wright, Wahba, PNAS August 2005. (Open Source)

N objects: Given a noisy possibly incomplete set of pairwise dissimilarity measures, d_{ij}. The class of Regularized Kernel Estimates (RKEs), are defined as solutions to optimization problems of the following form:

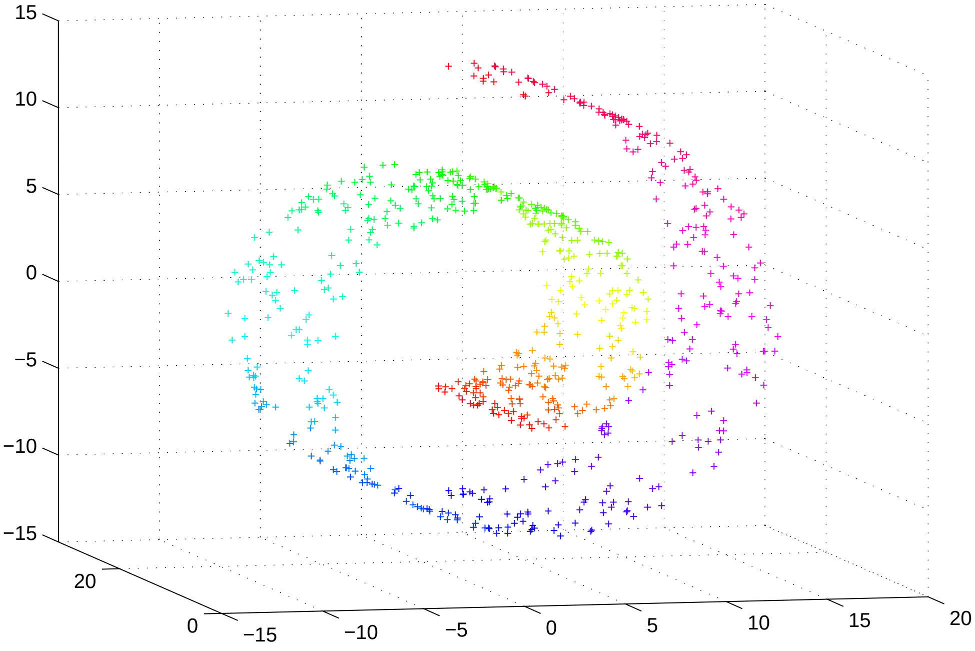
$$\min_{K \in S_N} \sum_{(i,j) \in \Omega} L\Big(d_{ij}, \hat{d}_{ij}(K)\Big) + \lambda J(K), \qquad (2)$$

S_N is the convex cone of all real nonnegative definite matrices of dimension N, Ω representative, (and connected) set of pairs, $\hat{d}_{ij}(K) = K(i,i) + K(j,j) - 2K(i,j)$ and $J(K) = trace K$. Solved using a convex cone algorithms (DSDP5, SDTT3). Choose λ, p. Pseudodata: Euclidean coordinates of ith object: $x(i) = (\sqrt{\lambda_1}\phi_1(i), \sqrt{\lambda_2}\phi_2(i), \cdots, \sqrt{\lambda_p}\phi_p(i)), i = 1, 2, \cdots, N$. (Eigenvalues and eigenfunctions of K) Newbie algorithm for embedding new objects is given.

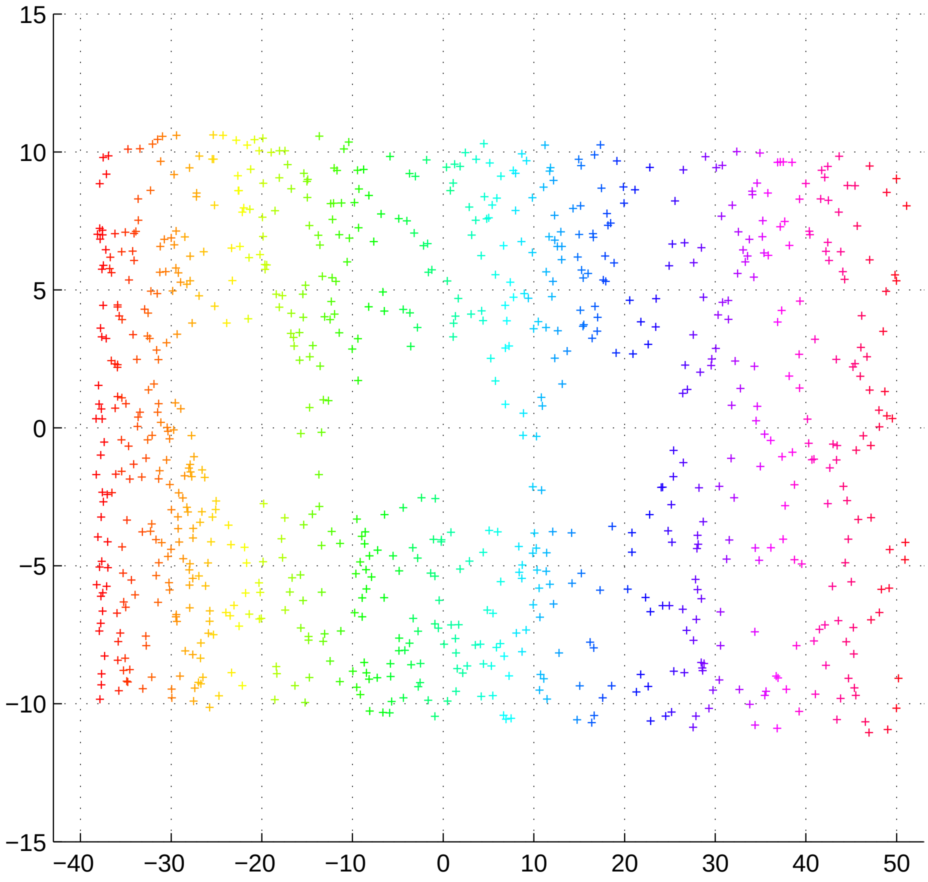
$L = \sum |d_{ij} - \hat{d}_{ij}|$, $J(K) = traceK$. Positioning test globin sequences in the coordinate system of 280 training sequences from the globin family. (Four subfamilies, α globins-red, β globins-blue, myoglobins-purple, “other” small subfamilies-green. Newbie algorithm used to locate one Hemoglobin zeta chain (black circle), one Hemoglobin theta chain (black star), seventeen Leghemoglobins (black triangles) into the coordinate system of the training globin sequence data.



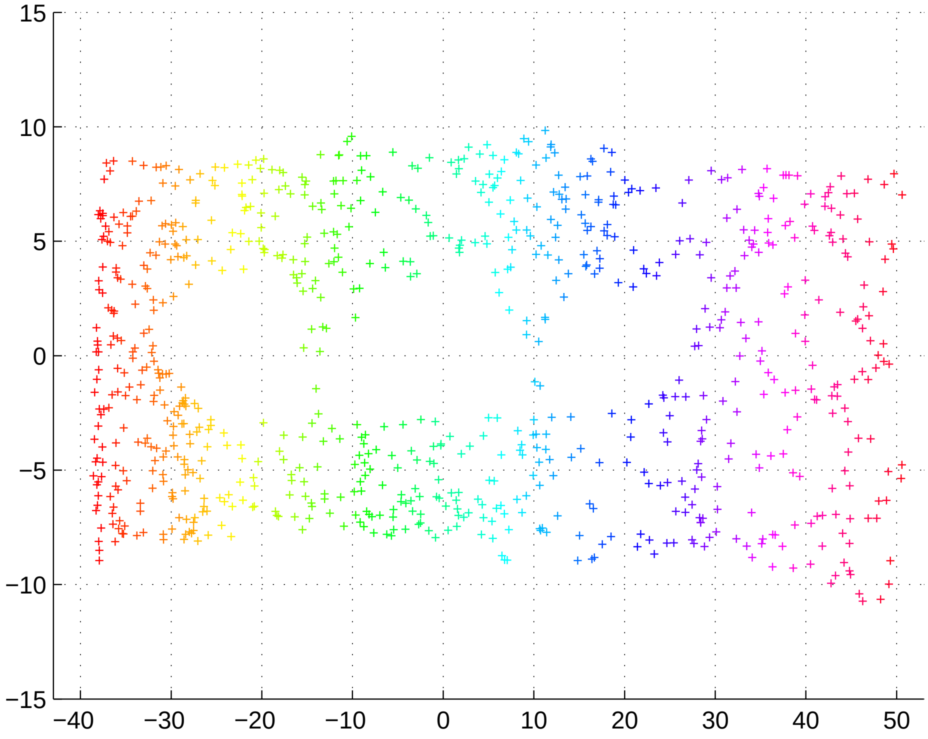
Robust Manifold Unfolding with Kernel Regularization Lu, Lin, Wahba, UW Statistics Dept TR 1108 October 2005. Same algorithm as PNAS Aug 2005 with the following inputs: Ω is now the set of k-nearest ij pairs, choose k, and $J(K) = -traceK$. This method will "unroll" the famous "Swiss Roll".



The Famous "Swiss Roll".



Wisconsin Roll, true parametrization.Observations come from rolled up version after adding noise-20% of pairwise observtions multiplied by a random number uniform between 0.85 and 1.15.



Noisy Wisconsin Roll, unrolled.