BREAST CANCER EPIDEMIOLOGY MODEL:

Calibrating Simulations via Optimization

Michael C. Ferris, Geng Deng,
Dennis G. Fryback, Vipat Kuruchittham

University of Wisconsin

University of Wisconsin Breast Cancer Simulation Model

Purpose –

- Use detailed individual-woman level discrete event simulation of processes
  - Breast cancer natural history
  - Breast cancer detection
  - Breast cancer treatment
  - Non-breast cancer mortality among US women
- To replicate historical breast cancer surveillance data, 1975-2000

How to screen even more!

ENHANCED AIRPORT SCREENING TO INCLUDE MAMMOGRAM

Early Detection Key to Cutting Down Terrorism, Women's Health Risks

San Francisco (SatireWire.com) — Arguing there is more than one way for a passenger to bring a "ticking time bomb" aboard an aircraft, the FAA today unveiled a new, more rigorous airport screening system that includes a mammogram.

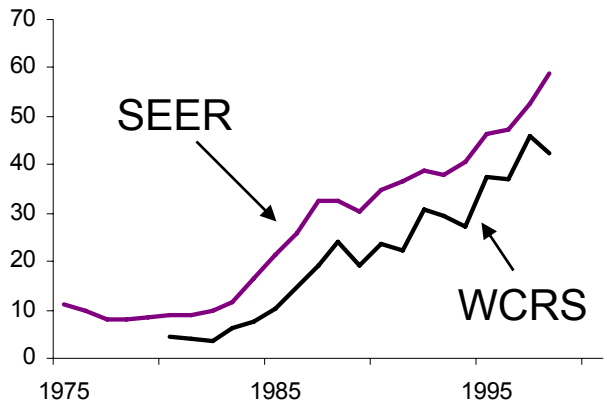
Although the new system has caused massive delays at airports where it has been tested, female passengers overwhelmingly approved of the added precautions.

"Sure, I worry that someone will bring a knife or a bomb on board, but what if I have a bomb slowly ticking away inside of me?" said 53-year-old Pamela Sardozian as she slowly disappeared with her luggage into an x-ray machine at San Francisco International Airport.

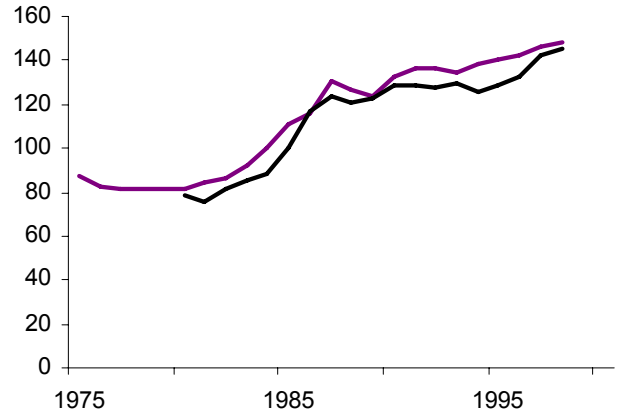
"Patienger" at O'Hare International

History of Breast Cancer Incidence 1975-2000

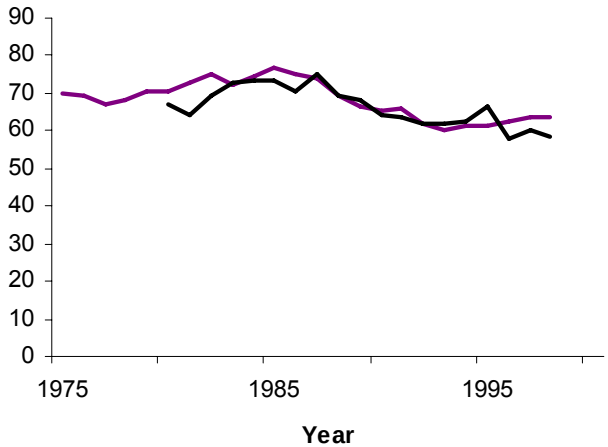
In Situ Inc./100K pop.



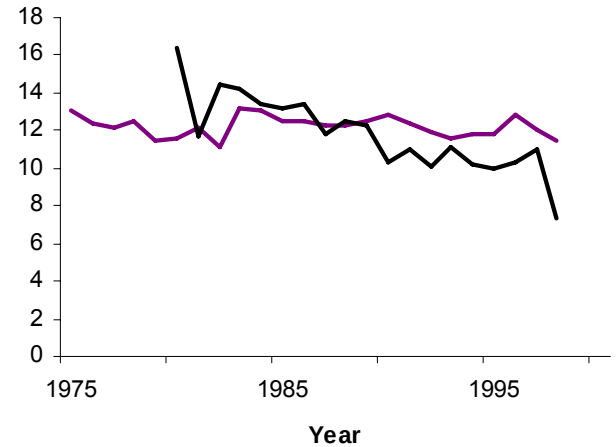
Localized Inc./100K pop.



Regional Inc./100K pop.



Distant Inc./100K pop.



Natural History

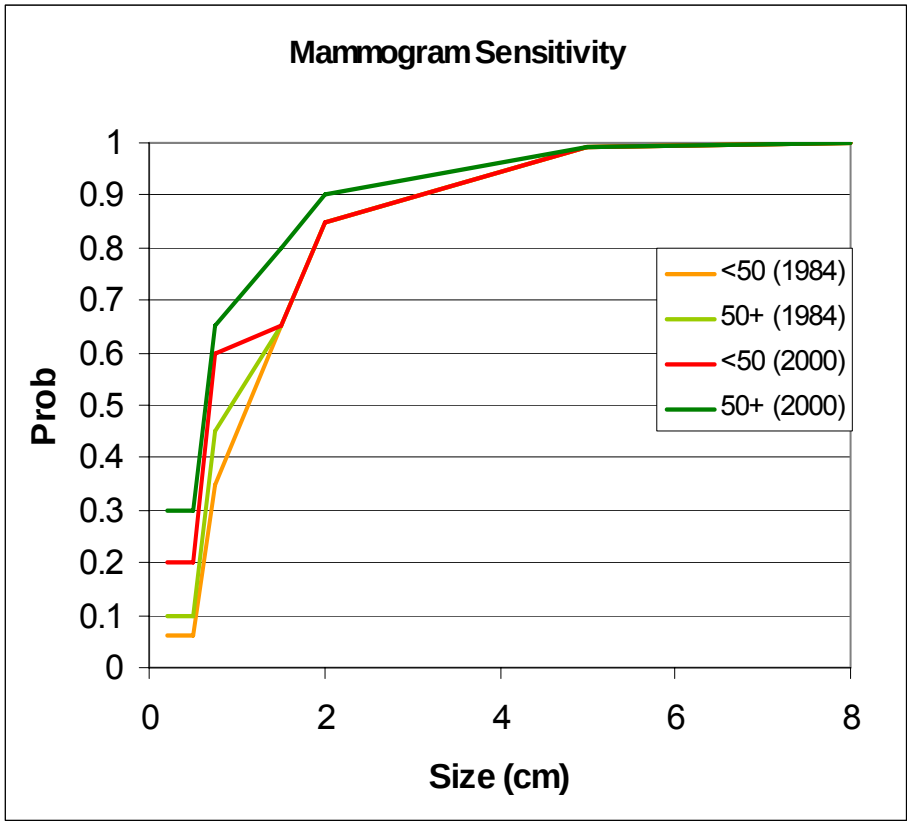
- Breast cancer occult onset rate, proportional to background incidence but lagged in time.

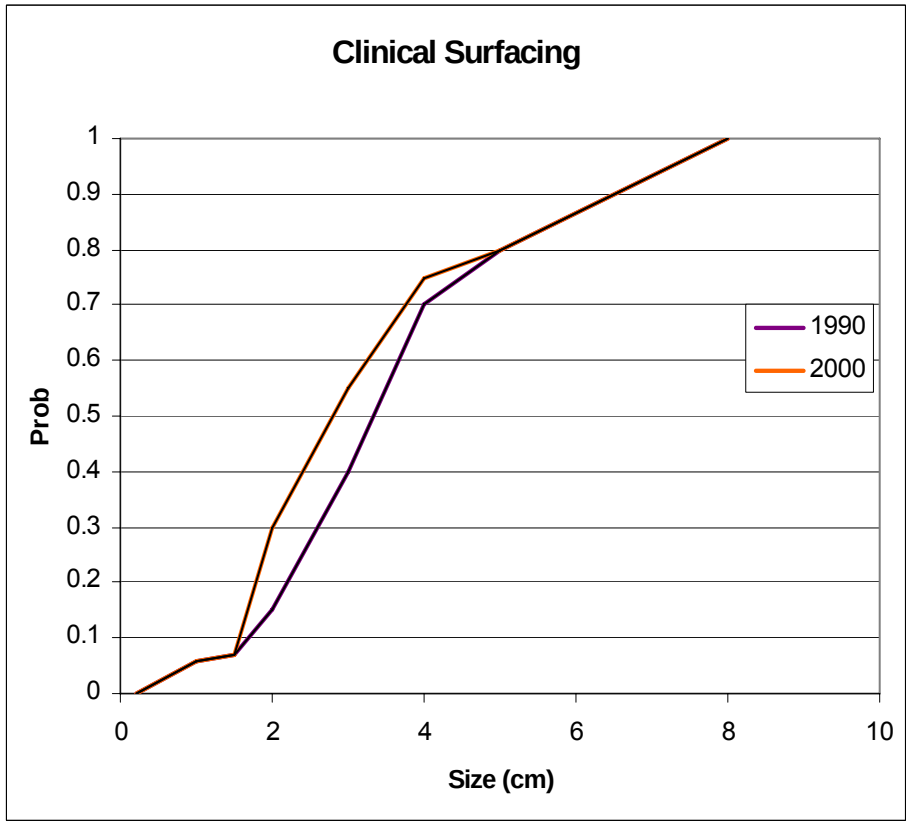
- Gompertzian growth from 2mm to 8 cm
  - Gamma distribution of initial growth rates
    - Fit mean and variance of this distribution

Breast Cancer Detection

- Screening sensitivity: probability of detection at screening as function of
  - Size of the tumor
  - Age of the woman (≤ 50 yr , > 50 yr)
  - Year in which she is being screened

- Non-screen detection (“clinical surfacing”)—annual probability as function of
  - Size of the tumor
  - Year (1975…2000)


Mammogram Sensitivity
Prob
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
8
Size (cm)
<50 (1984)
50+ (1984)
<50 (2000)
50+ (2000)


Clinical Surfacing
Prob
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
8
10
Size (cm)
1990
2000

Non-natural history parameters taken as fixed

- Mammography dissemination model
  - By year of simulation (1975-2000)
  - By age of simulated woman
- Breast cancer incidence in absence of screening.
  - Age-period-cohort model incidence fixed input
  - Calibrate occult onset to this
  - Calibrate an average lag in time between onset and incidence.
- Non-breast cancer mortality

The Problem

- Determine values for “free parameters” that make simulation closely track observations (2): optimize
- Develop a model of where “good parameters” are: approximate
- Develop biological insight by understanding model of good parameters: inference

Acceptance Sampling used to fit model parameters

1. Specify simulation model as function of randomly sampled free parameters.
2. Compute implied 25-year incidence & mortality curves
3. Determine fit to observed data using acceptance envelopes.

repeat this tens of thousands of times to find best fitting parameters (use large scale parallel computing to accomplish)

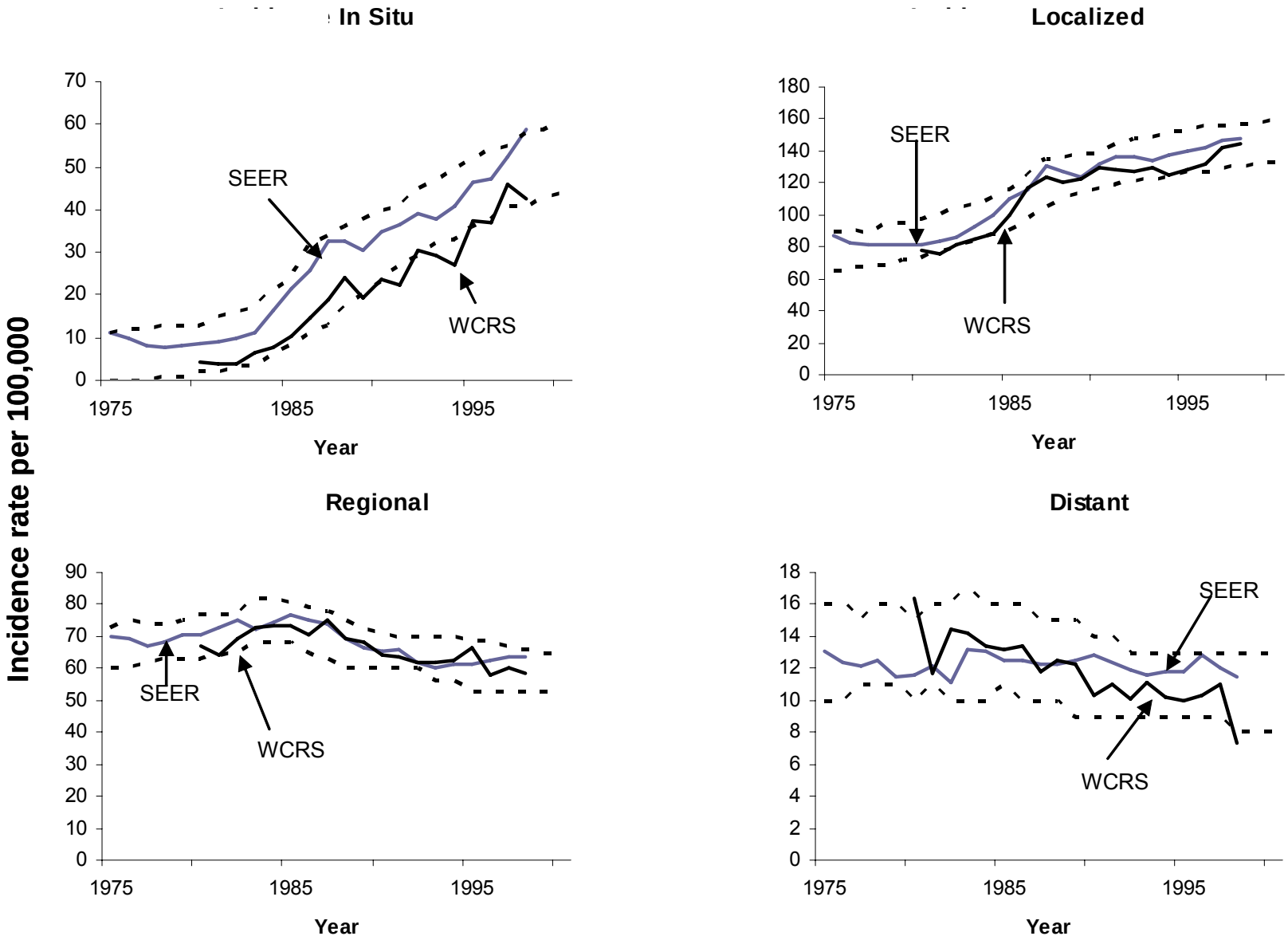
- Acceptance sampling also produces joint posterior distribution for all sampled parameters

Acceptance Envelopes

- Upper and lower bound placed around SEER and WCRS stage-specific incidence rate curves to screen for "acceptable" model-generated curves.

- Criteria for acceptance envelopes
  1. Envelops SEER rates and most of WCRS (biased toward SEER)
  2. Penalizes simulated curves not having inflections representing stage specific characteristics
     - *In situ* and localized stages must be increasing (no downturn)
     - Flattening of regional & distant stages over time
  3. Width set to encompass 95% variation expected in rates given size of population simulated

Envelopes to Score Incidence fit



Objective function

- *Given one replication with fixed input parameters*
  count of number of points at which model output falls outside of envelope
  - "worst" score = 104
  - "good" score ≤ 10
  - "best" score = 0
- Call this f(v)
- Really a discrete valued rank function
- Interested in "level sets" of function

$$L(\lambda) = \{v | f(v) \leq \lambda\}$$

Total sampled input parameters governing incidence of breast cancer:

- Natural history: 10
- Sensitivity of mammography: 4-16
- Annual surfacing: 2-12

…effectively something like 30 parameters

10 Parameters sampled

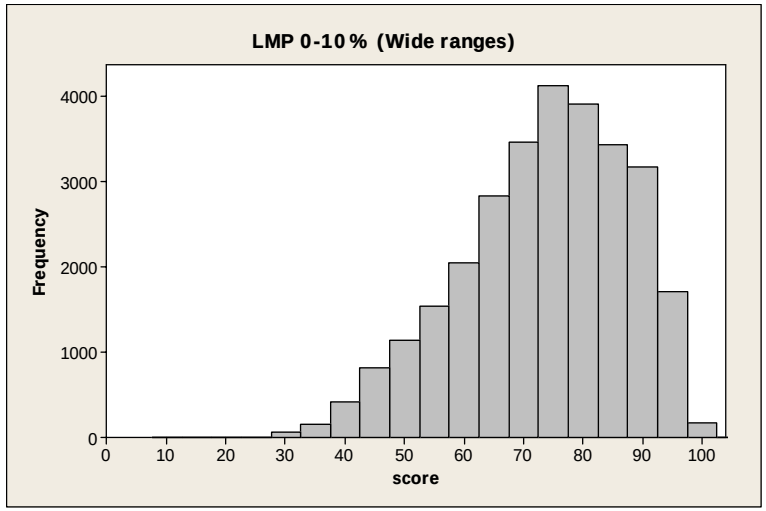
Parameter	Wide ranges	Focused ranges	Final value
1. Limited Malignant Potential (*LMP) Fraction*	[0% - 55%] (1%)	[30% - 50%] (1%)	42%
2. InSituBoundary	[0.75 – 1.0] (0.01)	[0.85 – 0.99] (0.01)	0.95 cm
3. *Max LMP Size*	[InSitu - 1.5 cm] (0.1)	1 cm	1 cm
4. *LMP Dwell Time*	[1-3] (0.5)	[1.5-2.5] (0.5)	2 y
5. *Onset Proportion*	0.85 – 1.2 (0.01)	0.8 – 1.0 (0.01)	0.9
6. *Onset Lag*	[1-8] (0.5)	[1.5-4] (0.5)	3 y
7. *Percent 4 nodes*	[0 – 5%] (1%)	[0 – 1%] (1%)	1%
8. *Percent 5 nodes*	[0 – 5%] (1%)	[2 – 4%] (1%)	2%
9. *Mean Gamma*	[0.01 – 0.2] (0.01)	[0.08 – 0.18] (0.01)	0.12
10. *Var Gamma*	[0.006 – 0.1] (0.001)	[0.01 – 0.05] (0.001)	0.012

Feasible Point Generation

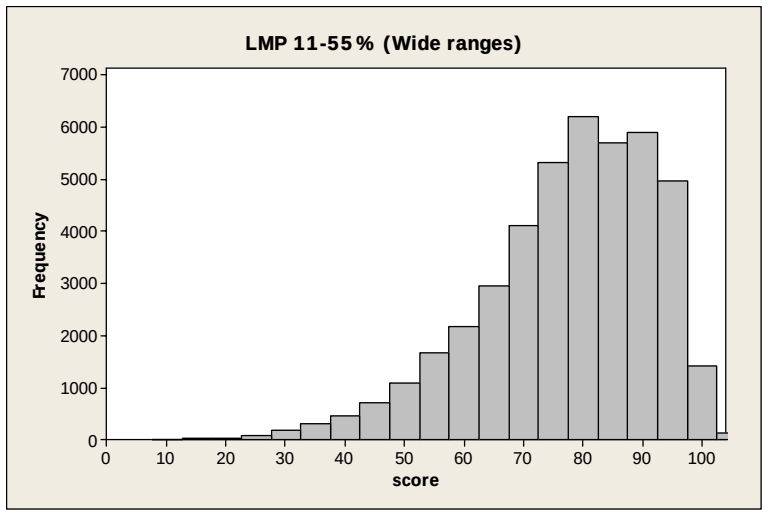
- 500,000 points (v) generated uniformly at random
- Using CONDOR (120 machines) can evaluate approximately 1000 per day
  - f(v) involves simulation of 3 million women
- 363 are feasible points (in L(10))
- Can we optimize?
- Can we characterize these points?

Finding Feasible Points

- First sample from wide ranges on all parameters in the model, but holding LMP% ≤ 10%. Result of 29,000 sampled input parameter vectors:

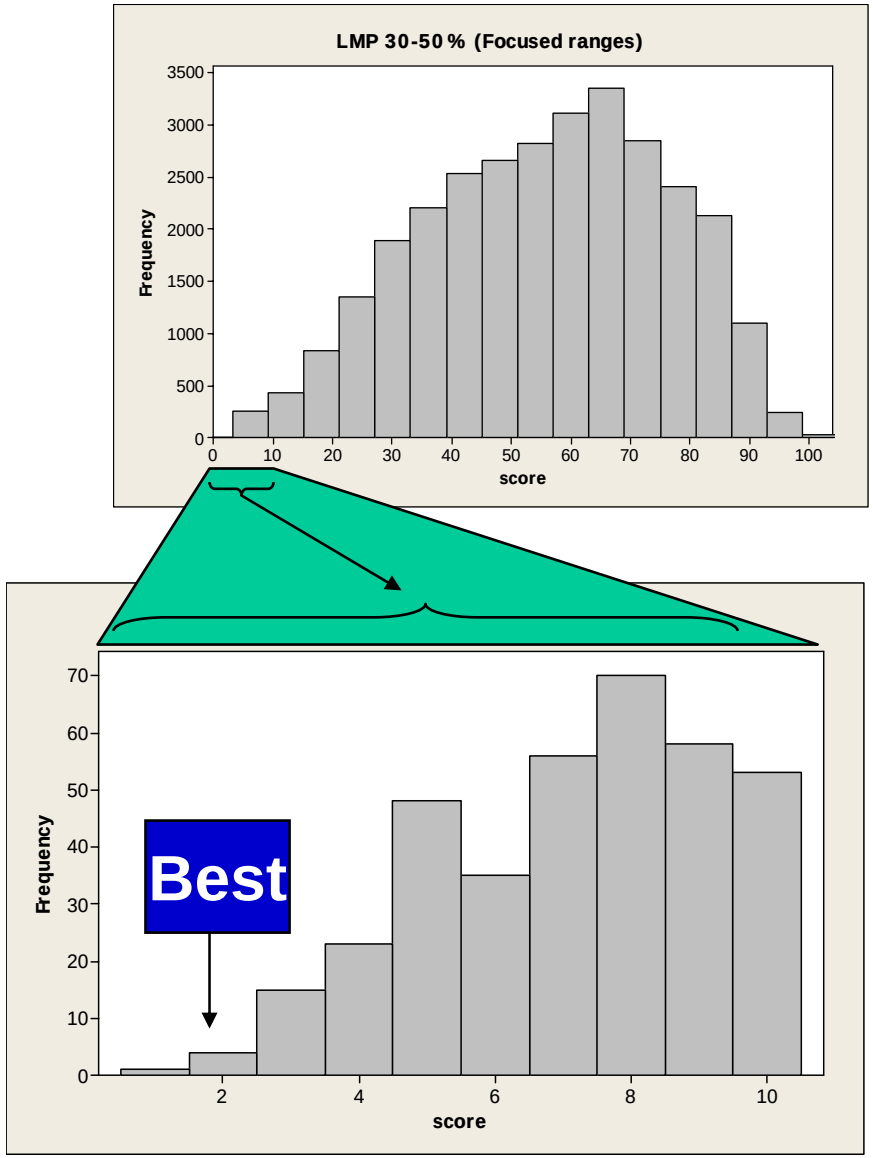


- Sample 43,000 vectors with 10%<LMP%<55%:



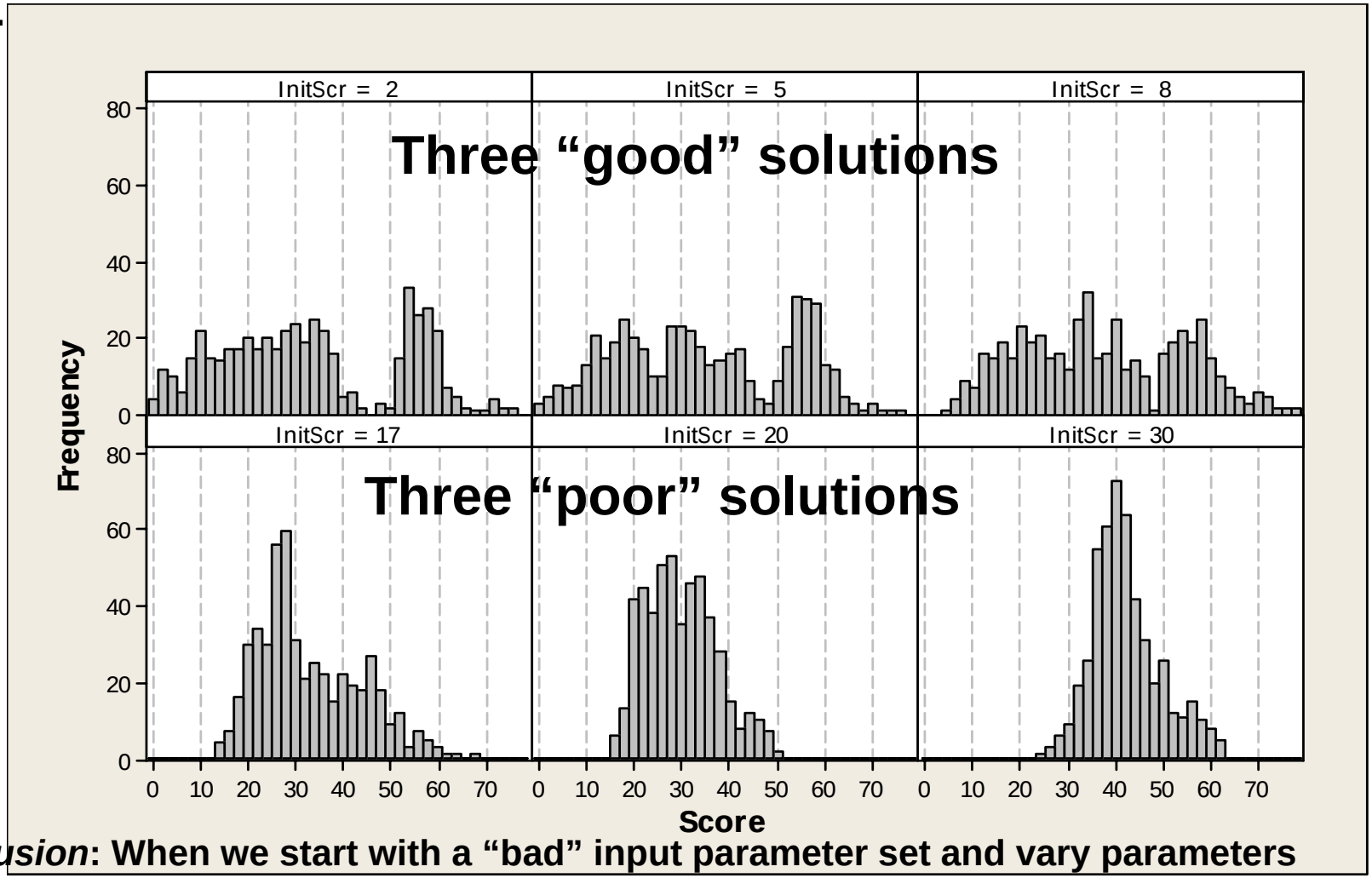
- *Conclude*: good solutions (score ≤ 10) are very rare
- *Conclude*: rule out LMP%<10%

- Now constrain 30%<LMP<50%, and hold other parameters in more focused ranges around our best solution and sample 30,000 new parameter vectors.
- Result is mostly poor solutions, but 363 parameter vectors have "good" scores.
- Conclude: good solutions are rare
- Question: do good solutions occur in relatively compact regions of parameter space?



Study 2) Neighborhoods around "Good" and "Poor" Solutions

Process: Pick starting input vector and sample 500 new input parameter sets in tight neighborhood around the starting one to see whether there is a good solution nearby.

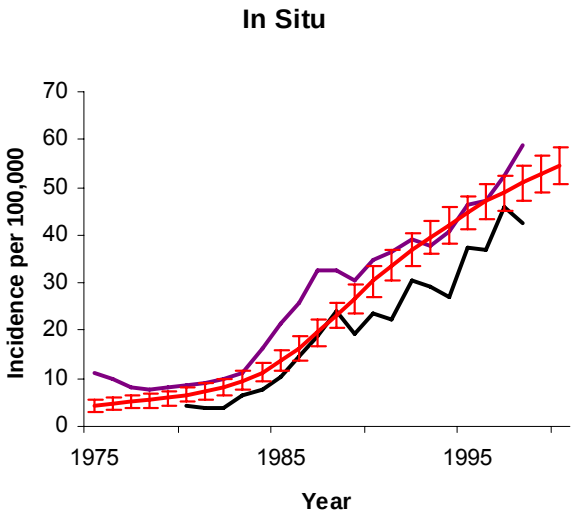


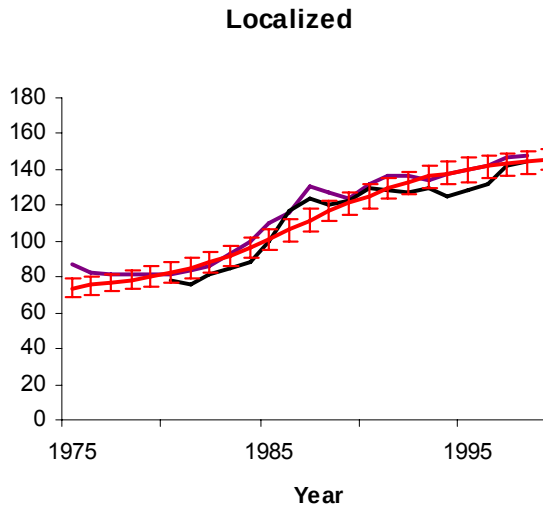
***Conclusion*: When we start with a "bad" input parameter set and vary parameters around it, we cannot find a good solution. Good solutions appear around good solutions, but are still rare.**

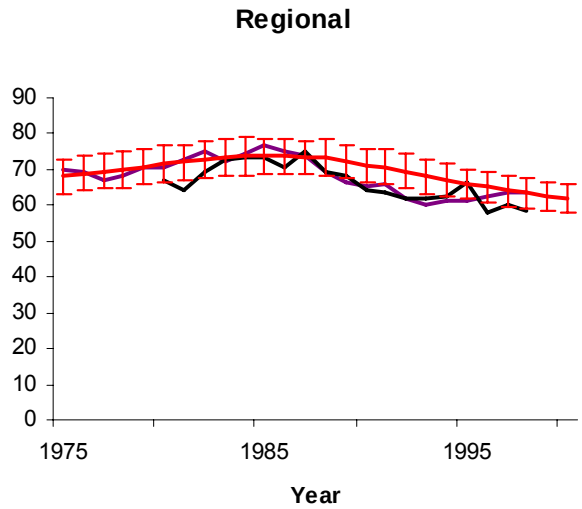
Stochastic Variation

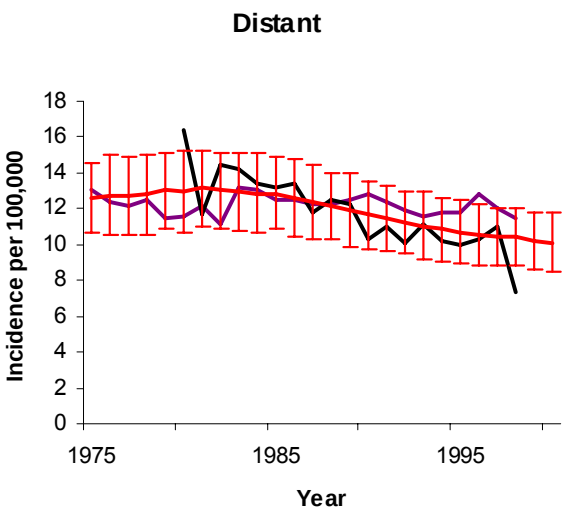
- Given fixed input parameters
  there is still stochastic variation
  of breast cancer incidence and mortality
- Each replication is 1 alternative "history"
  of breast cancer incidence and mortality
  over years 1975-2000 for population
  size of Wisconsin
- Score each replication $f_\omega(v)$
- Multiple replications reduce feasible pts
- Get histogram of scores

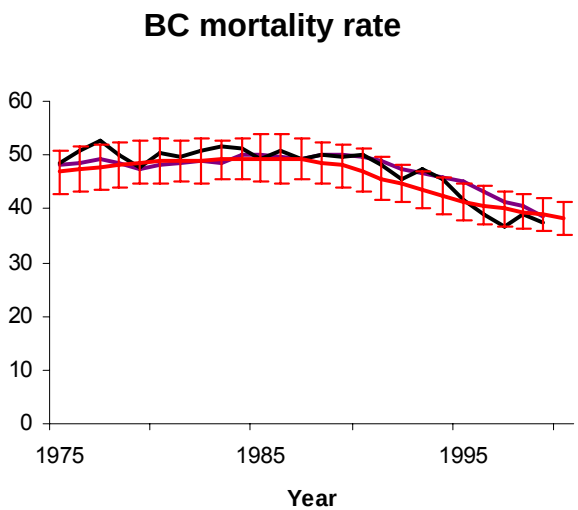
300 replications of best model (95% intervals)

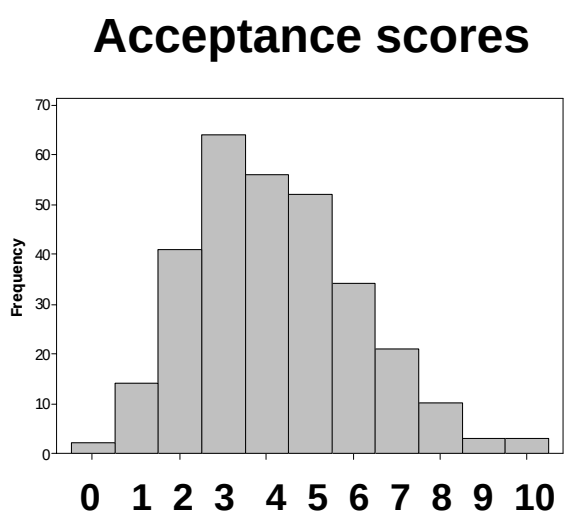








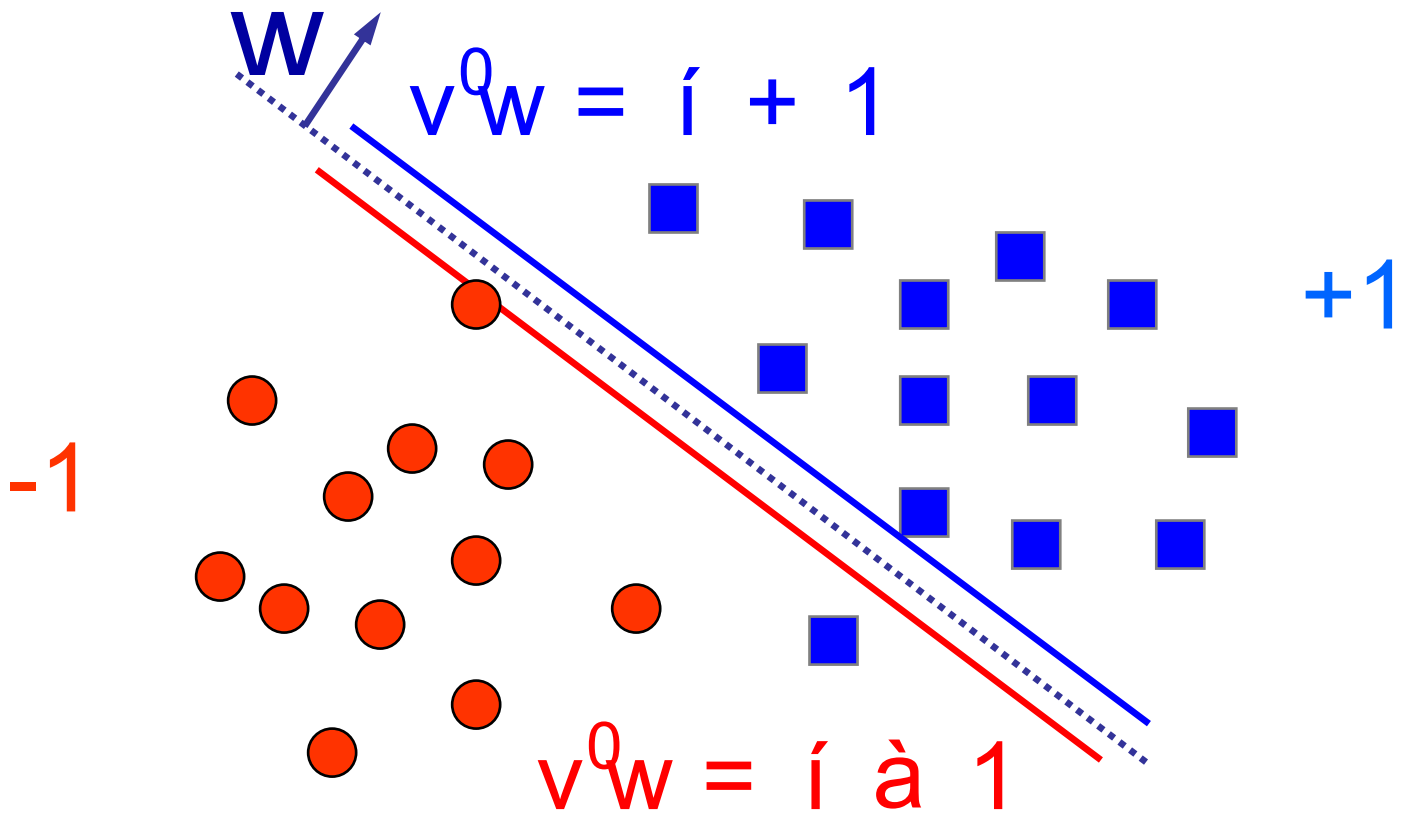




Process

- Split data into training and testing set
- Solve optimization problem over training set to generate classifier
- Validate using testing set
- Classifers:
  - SVM
  - kNN (nearest neighbour)
  - C45 (decision tree)
  - Boosting (multiple classifiers then vote)
- Only simulate points classified as good

Linear Classifier



Separation Problem

- P_+ and P_- are two populations
- $A_+ \in \Re^{m_1 \times k}$ and $A_- \in \Re^{m_2 \times k}$ measure characteristics
  - m_1 and m_2 – number of samples
  - k – number of features measured per sample
  - $m_1 + m_2 \gg k$
- Separate populations with hyperplane: $\{x \mid x^T w = \gamma\}$

$$A_+ w > 1\gamma$$
$$A_- w < 1\gamma$$

- Normalize

$$\begin{aligned} A_+ w - 1\gamma &\geq 1 \\ A_- w - 1\gamma &\leq -1 \end{aligned}$$

Misclassification Minimization

- Let D be a diagonal matrix

$$D_{i,i} = \begin{cases} 1 & \text{if } i \in P_+ \\ -1 & \text{if } i \in P_- \end{cases}$$

- Separation condition

$$D(Aw - 1\gamma) \geq 1$$

- Generally problems are not separable

- Minimize misclassification error

$$\begin{aligned} &\min_{w,\gamma,y} && \tfrac{1}{2} \|y\|_2^2 \\ &\text{subject to} && D(Aw - 1\gamma) + y \geq 1 \end{aligned}$$

Linear Support Vector Machine

- **Select one with maximum separation margin**
- **Example formulation**

$$\begin{aligned} \min_{w,\gamma,y} \quad & \frac{1}{2}\|w\|_2^2 + \frac{\nu}{2}\|y\|_2^2 \\ \text{subject to} \quad & D(Aw - 1\gamma) + y \geq 1 \end{aligned}$$

  - $\frac{2}{\|w\|_2}$ – separation margin
  - $\|y\|_2^2$ – misclassification error
  - ν – weighting of the goals

Wolfe Dual

$$\begin{aligned} \min_{x} \quad & \tfrac{1}{2\nu} x^T x + \tfrac{1}{2} x^T D A A^T D^T x - e^T x \\ \text{subject to} \quad & e^T D^T x = 0 \\ & x \geq 0 \end{aligned}$$

- Strongly convex quadratic program
- Quadratic term is a rank-k update to a pd matrix
- Hyperplane recovered

  $w = A^T D^T x$

  $\gamma =$ multiplier on $e^T D^T x = 0$

Classifier

$$f(v) = v'w - \gamma$$

$$f(v) > 0 \Rightarrow v \in P_+$$

$$f(v) = v'A'u - \gamma$$

$$f(v) = K(v', A')u - \gamma$$

- Kernel generates nonlinear separator
- Linear, Polynomial, Gaussian

Nonlinear Support Vector Machine

$$\min_{u,y \geq 0} \quad \|u\| + \nu \|y\|$$
$$\text{subject to} \quad D(K(A, A')u - \gamma \mathbf{1}) + y \geq \mathbf{1}$$

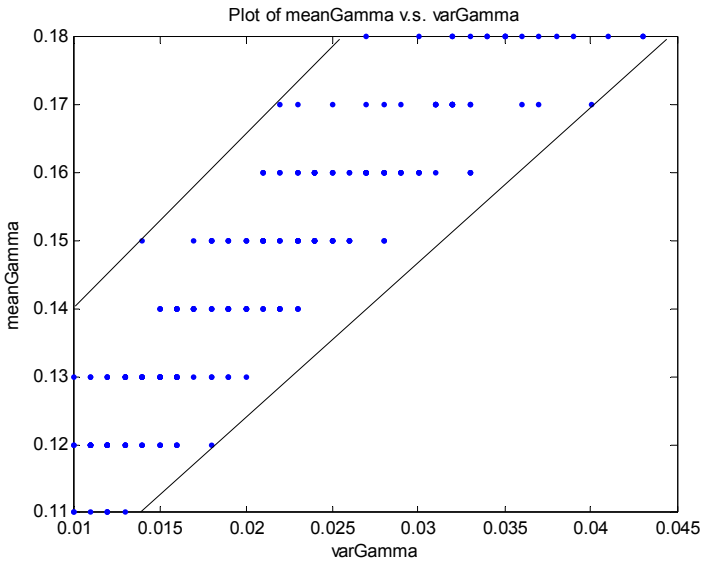
- Tradeoff errors and margin
- Can interpret this as columns of A' indexing basis functions (reduced SVM)
- Feature selection removes columns of A; can be formulated as a (non)linear MIP
- Large scale (dense) optimization

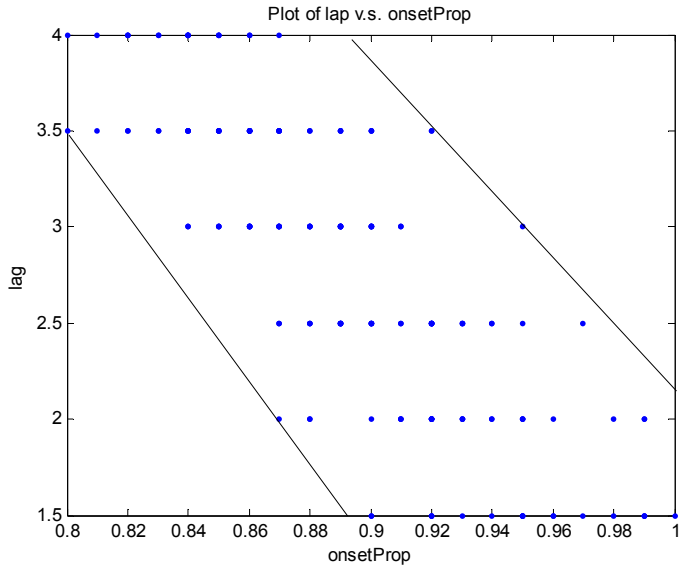
Classifier Evaluation

- Imbalanced data; many more negative points
- Accuracy is great by classifying all points as negative!
- More effective to use TP and TN
- Can generate classifers with blue or red validations

actual		predicted +	predicted -
	+	TP 0.95 0.6	FN 1
	-	FP	TN 1 0.4 0.7

Cutting planes (2d projections)

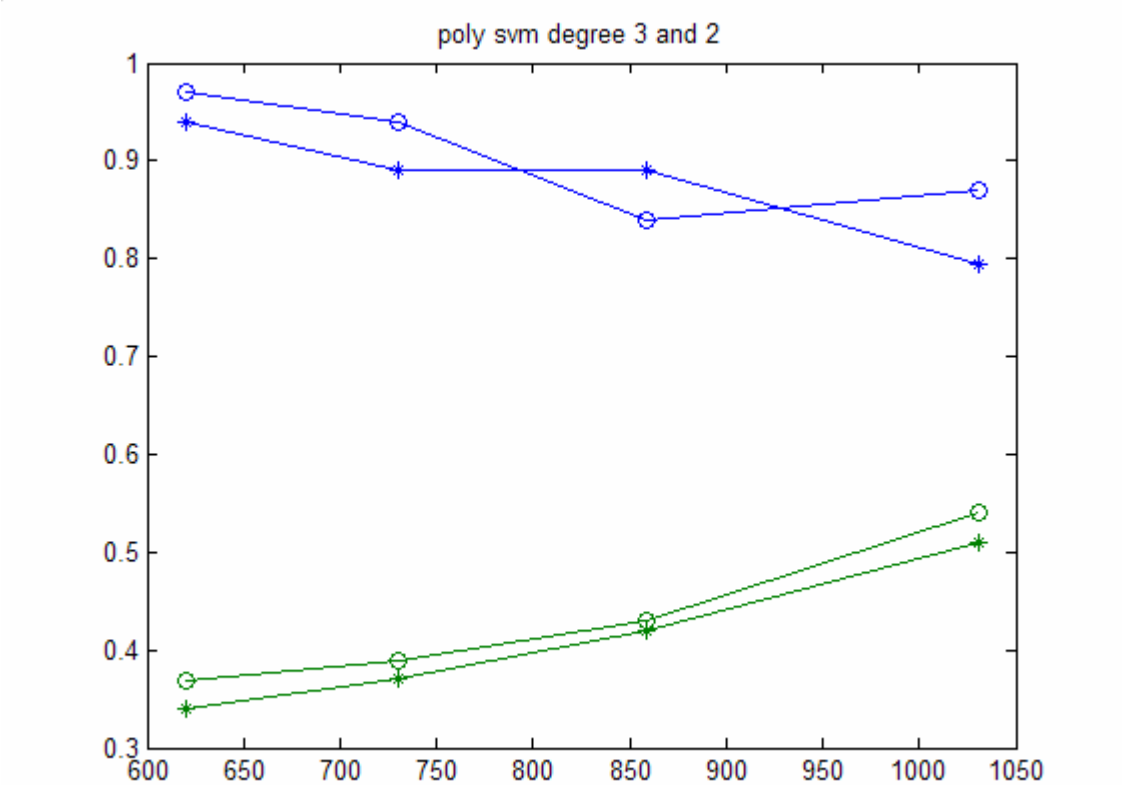




Training/Testing set refinement

- |TP0| = 320, |TN0| large!
- One-sided sampling (Kubat, Matwin)
  - Generate consistent subset of TN0 (C) using 1NN (from random initial set)
  - Recursively perform “Tomek” reduction until |C| is suitable
- Resampling with replacement
  - Increase size of TP0 using “score” weighted probablilites

Accuracy (TP/TN)



Generation method

- Generate 100,000 samples uniformly at random (potential values for v)
- Use naïve cutting planes from data to remove very poor samples **8640 remaining**
- Sequentially generate classifiers using one sided sampling with high estimated TP
- Remove negative points from sample
  - Some positive points removed **788 remaining**
  - Many more negative points removed

Reality check

- Bad news – using “experts” values for remaining 20 parameters only 20% of these evaluate as P+
- Good news – using Bayesian estimate with 363 “+” as a prior for remaining 20 parameters gives 65% as P+
- Experts agree that Bayes estimate values are good

Results (II)

- Run one more classifier that has TP accuracy of .4 and TN accuracy of .9
- Maybe sacrifice some “+”’s
- Resulting 220 points – 195 are P+

- Open research: do feature selection to determine which of the 30 parameters are the most important

Stochastic function values

- New dataset with 10 replications at points with scores ≤ 30
- Update classifiers to use replication data
- Far fewer points in TP
- Generation process results in new points (all are good), but 2 of which seem better than the “experts best solution”

- Utilize derivative free optimization code or response surface methodology to really optimize

Surrogate optimization

- Use “Dace” toolbox to generate interpolant
- Smoothness dependent on number of “training values” (2400, use max)
- Use Nelder Mead simplex method, or nonlinear linesearch method for optimization of this surrogate
- New point generated evaluated 300 times

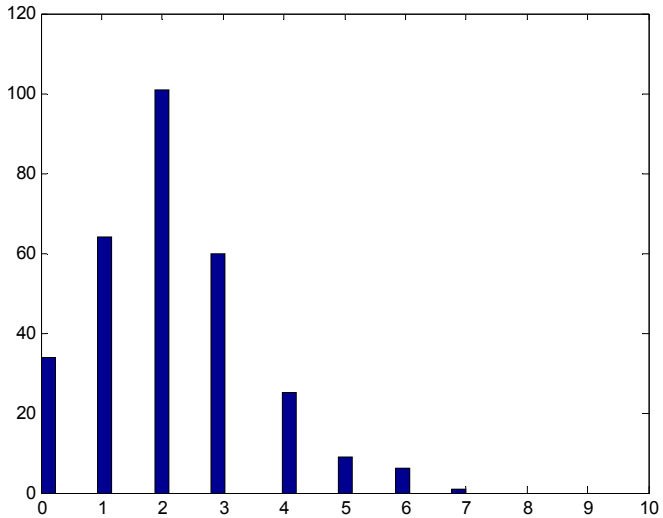
What obj values for DACE?

Define objective as	Average of max	Average of mean
(trained by max of reps)	5.25	3.62
(trained by mean of reps)	5.45	3.17
(trained by max + 0.1 mean)	5.35	3.42
(the first function is trained by max and the second is trained by mean)	5.37	3.43

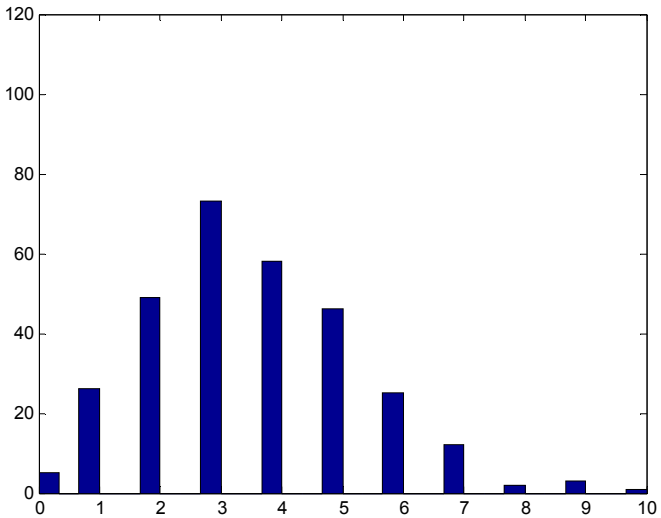
Function domain/grid

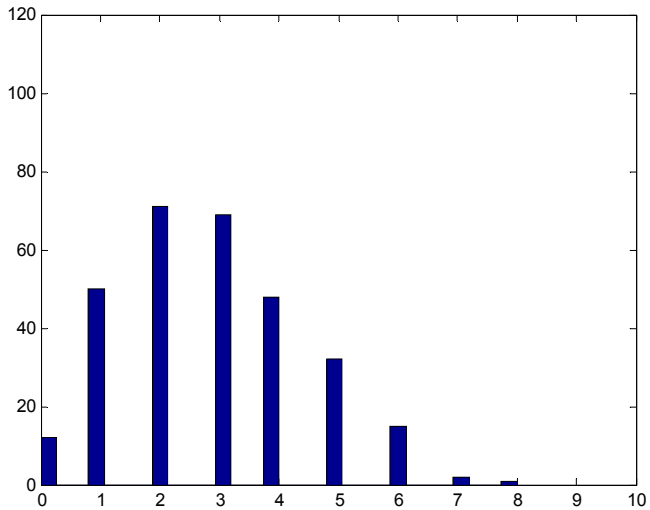
	Min	Max	Min grid size	range
onsetProp	0.8	1	0.01	0.2
meanGamma	0.1	0.18	0.01	0.08
varGamma	0.01	0.045	0.001	0.035
boundInsitu	0.85	0.99	0.01	0.14
LMPFract	0.3	0.5	0.01	0.2
aggr4Node	0	0.01	0.01	0.01
Aggr5Node	0.02	0.04	0.01	0.02
lag	1.5	4	0.5	2.5
LMPRegress	1.5	2.5	0.5	1.0

Point/refinement tradeoff



2400 points





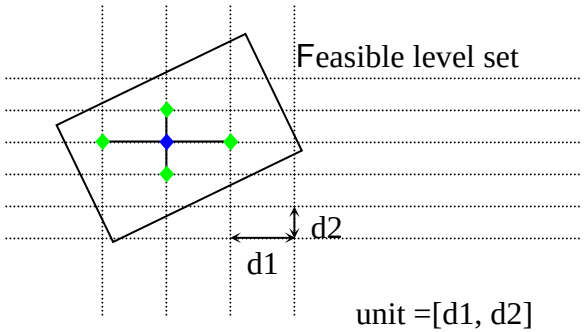
(1 refine) 1500 points (19 refine)

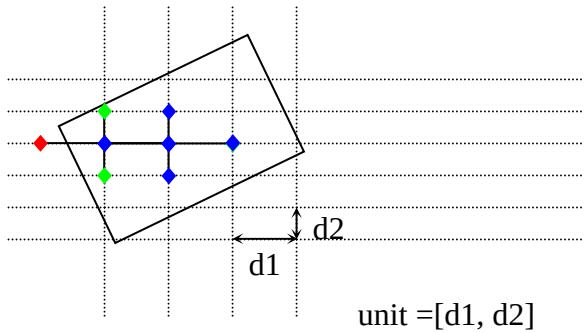
Representation

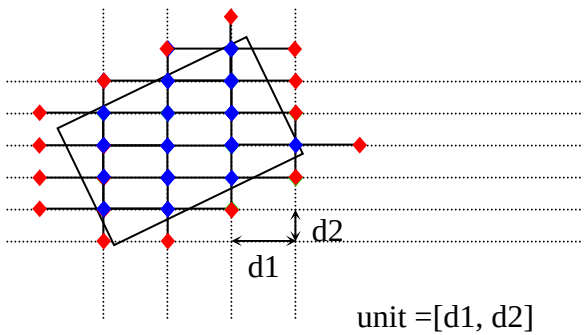
- Describe feasible set:

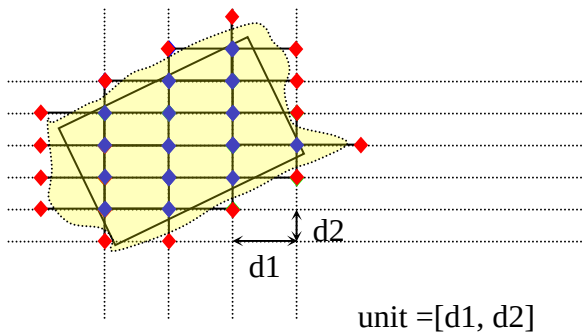
$$\{v \mid \max_{\omega} f_{\omega}(v) \leq 10\}$$

- Can generate more members of set
- Set is not connected – understand biological significance of each piece
- Does a representation of feasible set lead to fuller understanding of flaws in the simulation


Feasible level set
d2
d1
unit =[d1, d2]


d2
d1
unit =[d1, d2]


d2
d1
unit =[d1, d2]


d2
d1
unit =[d1, d2]

Island quality

Island 1	97/98 points have values <=10
Island 2	20/49 points have scores <= 10
Island 3	97/107 points have scores <= 10
Island 4	59/69 points have scores <= 10
Island 5	59/61 points have scores <= 10

Conclusions

- Effectively generate parameter settings for complex simulations
- Feasible sets may form disconnected "islands" in parameter space indicating possible biological switches governing biological behavior of the system or other complex system behaviors of interest.
- Optimization improves understanding; imbalanced datasets are widespread

Key issues

- Stochastic (expensive) function values
- Derivatives not (easily) computable

- Classifiers applied to unequally sized data sets

- User defined objective without smoothness/connectivity properties
- Approximation in “good” regions

Implications

- In 2000
  - 44% of in situ breast carcinomas diagnosed were LMP
  - 31% of small localized invasive breast cancers were LMP
  - 30% of all breast cancer survivors alive were LMP
- These are women who are being treated for a pseudo disease
- If substantiated, need work to find biological marker for LMP to avoid over treating