

# Genotype x Environment Interaction in Plant Breeding: To Avoid It or To Exploit It?

Rong-Cai Yang<sup>1,2</sup>

<sup>1</sup>Alberta Agriculture and Forestry

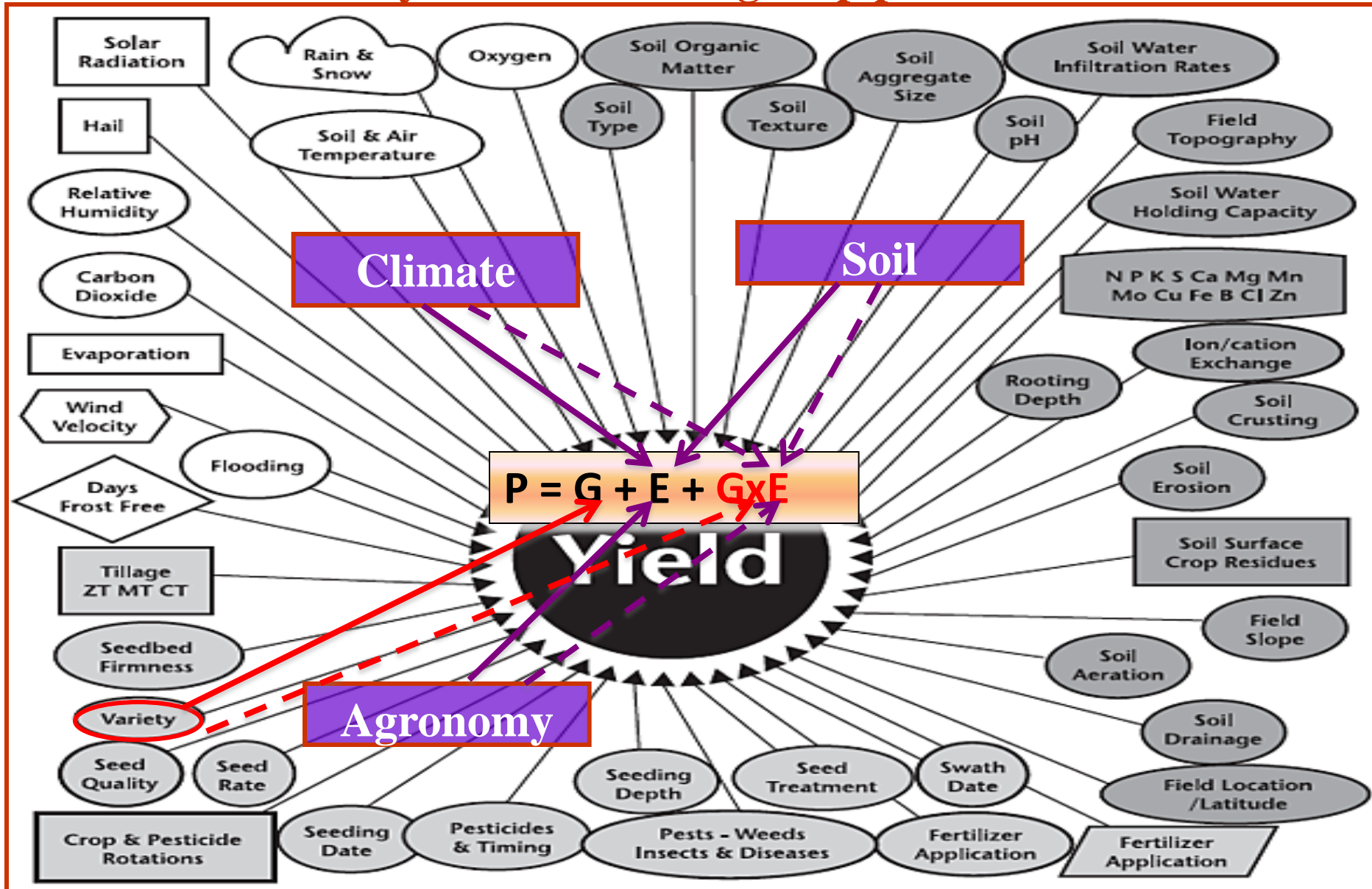
<sup>2</sup>University of Alberta

June 24, 2015

# What is GxE?

- When comparing responses of two genotypes to environmental gradients, **GxE is the failure of the two response curves to be parallel** (Baker 1988, Proceedings of ICQG 1987).
- Similar concept in evolution and ecology but often different terms are used: **phenotypic plasticity (robustness)**, **reaction norm** (e.g., Via 1994)

# Big challenge to characterize environments...when there are so many factors affecting crop production!!



# Designs for studying GxE

- **Agricultural crops:**

- Inbred lines can be self-pollinated to produce large numbers of genetically identical progeny
- Same genotypes can be replicated over environments
- Multiple-environment trials (METs)

- **Forest trees**

- Not possible to replicate same genotypes over environments
- 'Replication' over environments is realized by using relatives from the same family (e.g., half-sibs)
- Provenance trials or provenance-progeny trials.

# Barley cultivar trials across Canadian Prairies:

**GxE is an important component of total variability**

Year	Source	Alberta		Saskatchewan		Manitoba	
		df	% SS	df	% SS	df	% SS
1995	E	15	77.95				
	G	38	6.52				
	G × E	570	15.53				
1996	E	19	83.29				
	G	43	7.08				
	G × E	690	9.63				
1997	E	17	57.49				
	G	38	20.39				
	G × E	623	22.12				
1998	E	18	81.14				
	G	38	8.57				
	G × E	684	10.29				
1999	E	19	85.12	14	67.44	14	85.89
	G	38	5.73	35	12.92	33	5.10
	G × E	656	9.14	490	19.64	238	9.00
2000	E	21	71.45	22	66.06	12	84.49
	G	38	4.68	39	18.66	28	4.95
	G × E	750	23.87	858	15.27	132	10.56
2001	E	19	79.71	19	75.09	11	82.96
	G	34	7.61	37	12.88	19	8.44
	G × E	553	12.68	627	12.03	121	8.60
2002	E	15	68.54	13	87.91	14	84.64
	G	41	9.97	37	5.59	23	5.58
	G × E	458	21.49	412	6.49	211	9.78
2003	E	18	89.69	12	95.63	14	82.70
	G	40	3.71	21	0.80	17	5.25
	G × E	569	6.60	221	3.57	162	12.05

- **GxE > G in most (17/19) cases**
- **G <10% in most (16/19) cases**
- **GxE is up to 24%**

From Yang et al. (2006)

# GxE: a challenge or an opportunity?

- **Avoid it:** GxE has a large *unpredictable* component (e.g., due to year-to-year weather fluctuations). Select for lines which are more stable over all environments to mitigate GxE effects
- **Exploit it:** GxE has a large *predictable* component (cultural practices/adaphic/climatic patterns). For example, select for genotypes that are highly responsive to cultural practices (e.g., to crop inputs such as fertilizers)

# Make use of multiple-environment trials (METs) to help breeders to cope with GxE

- GxE interaction in MET can be decomposed into contributions for locations, years, and location x year interactions,

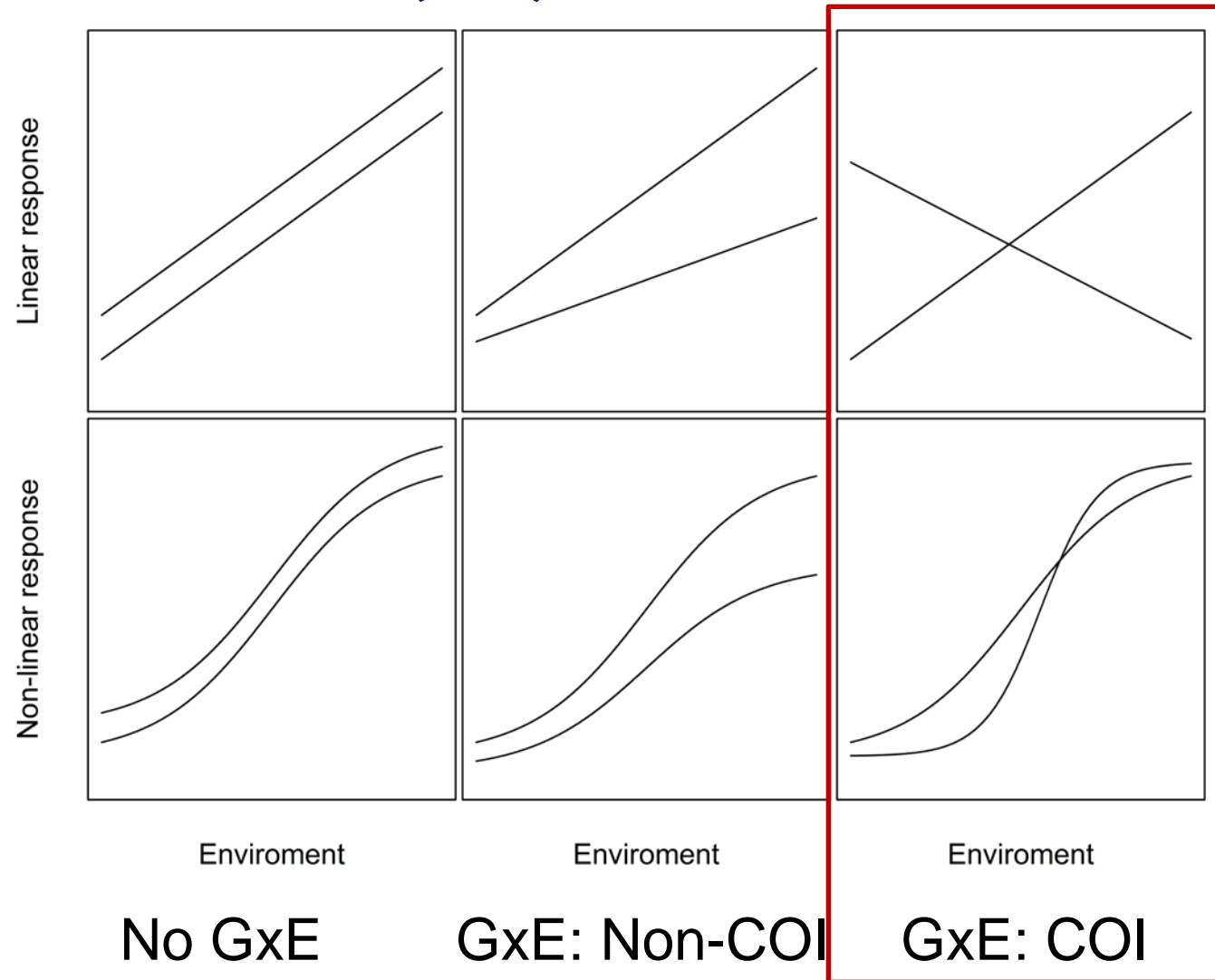
$$V(\text{GxE}) = V(\text{GxL}) + V(\text{GxY}) + V(\text{GxLxY})$$

- The ratio,  $V(\text{GxL}) / V(\text{GxE})$ , is the relative contribution to the GxE variance by predictable environmental factors
  - If the ratio is small, then **avoid GxE** → breeding for stability
  - If the ratio is large, then **exploit GxE** → breeding for location

From Walsh & Lynch Vol. 2

([http://nitro.biosci.arizona.edu/zbook/NewVolume\\_2/newvol2.html#2B](http://nitro.biosci.arizona.edu/zbook/NewVolume_2/newvol2.html#2B))

# Identifying unpredictable GxE: Crossover GxE (COI)



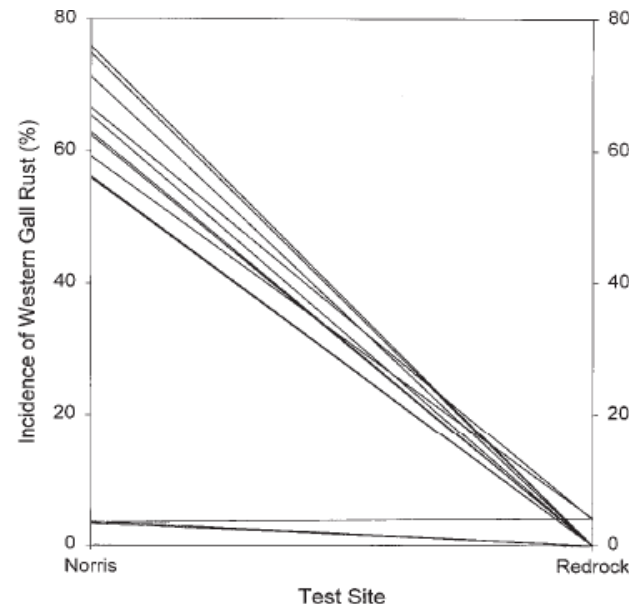
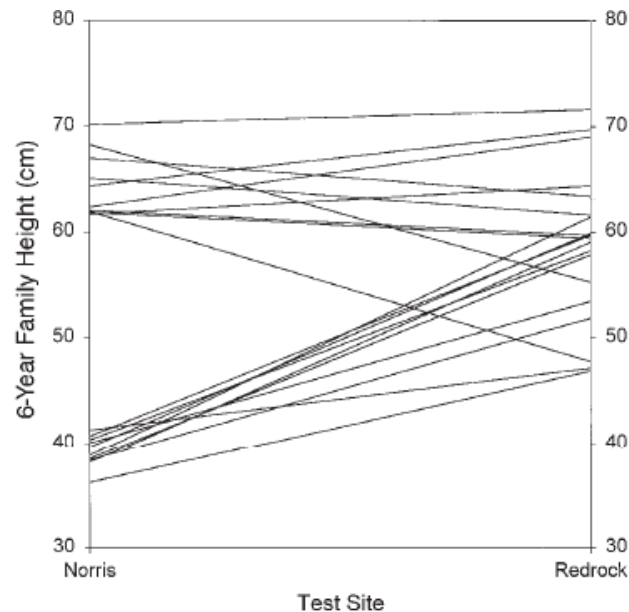


# COI GxE is ubiquitous, but present only with the presence of varying environmental conditions

## Comparative assessment of genetic variation of young high-elevation lodgepole pine for height and western gall rust resistance across two sites in Alberta CJFR 28:478-484 (1998)

Rong-Cai Yang, Narinder K. Dhir, and Leonard K. Barnhardt

- Select 10 best and 10 worst families at Norris site, but COI occurred when looking at both sites
- GxE for stress-related traits (e.g., WGR) can be assessed only in the presence of stress

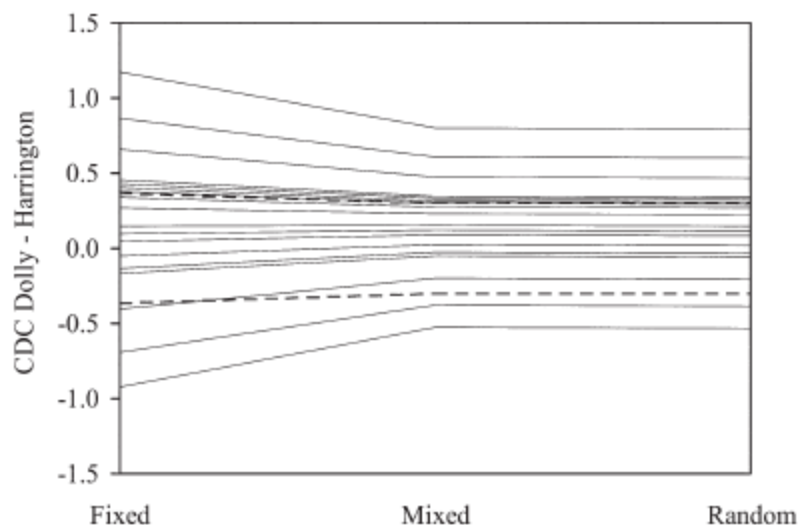


# Modified Azzalini–Cox test for COI GxE

## Mixed-Model Analysis of Crossover Genotype–Environment Interactions

Crop Sci 47:1051-1062 (2007)

Rong-Cai Yang\*



- $g(g-1)e(e-1)/4$  possible quadruples [e.g., 2,295 quadruples for 6 genotypes and 18 environments]
- Interaction-wise error rate  $(\alpha/2)^{0.5} = 0.158$  rather than comparison-wise ( $\alpha = 0.05$ ) or even more conservative experiment-wise (the original Azzalini-Cox test) error rate
- The test for COI is more conservative in mixed and random models than in fixed model due to shrinkage of random GxE effects

**The original Azzalini–Cox test gives experiment-wise error rate protection against rejecting a true null hypothesis (lower Type I error rate) at a cost of high Type II error rate (i.e., low power to detect the true COI). However, a Type I error may not be serious because follow-up cultivar trials will reveal spurious COI, but a Type II error is serious because a potentially important COI may go undetected.**

# Dealing with unpredictable GxE: stability analysis

- Linear approach: classical stability analysis
- Linear-bilinear approach: Seeking structured patterns in GxE
- Nonlinear approach: Capturing more GxE variation

# Classical stability analysis pioneered by Yates & Cochran (1938)

- Create a single environmental index as a surrogate to represent all of the complex and unobservable characteristics of the environment
- Carry out a simple linear regression analysis:

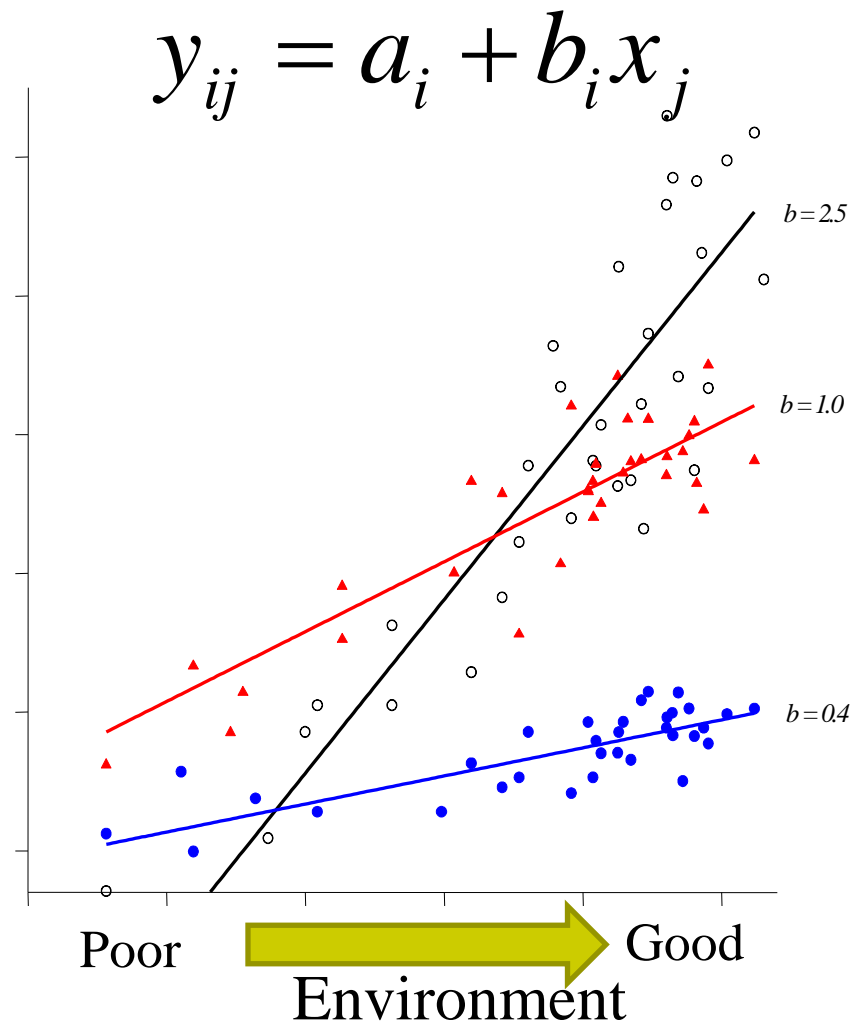
$$y_{ij} = a_i + b_i x_j$$

where  $y_{ij}$  is the performance (yield) of the  $i$ th genotype tested in  $j$ th environment;  $x_j$  is the mean yield of all genotypes tested in the  $j$ th environment (**environmental index**)

# Interpreting classical stability analysis in plant breeding context (e.g., Finlay & Wilkinson 1963, Eberhart & Russell 1966; Perkins & Jinks 1968)

Stability of a genotype = regression coefficient :

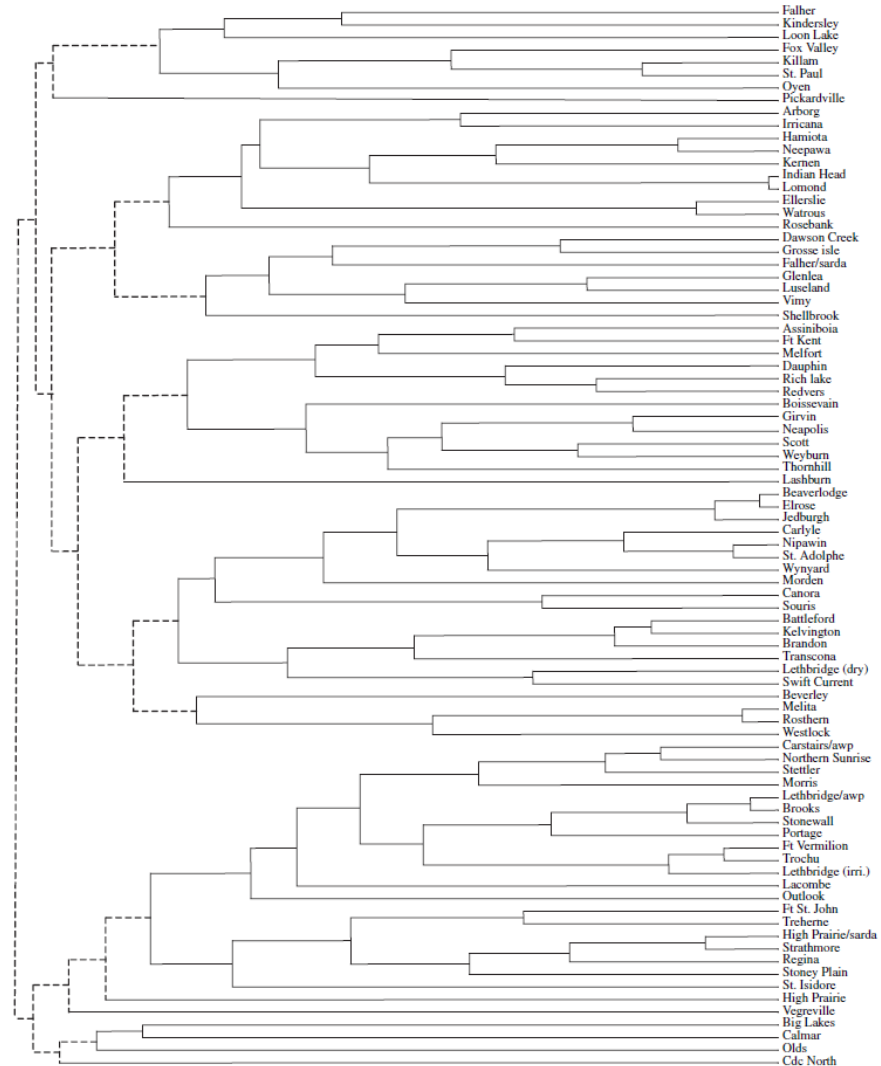
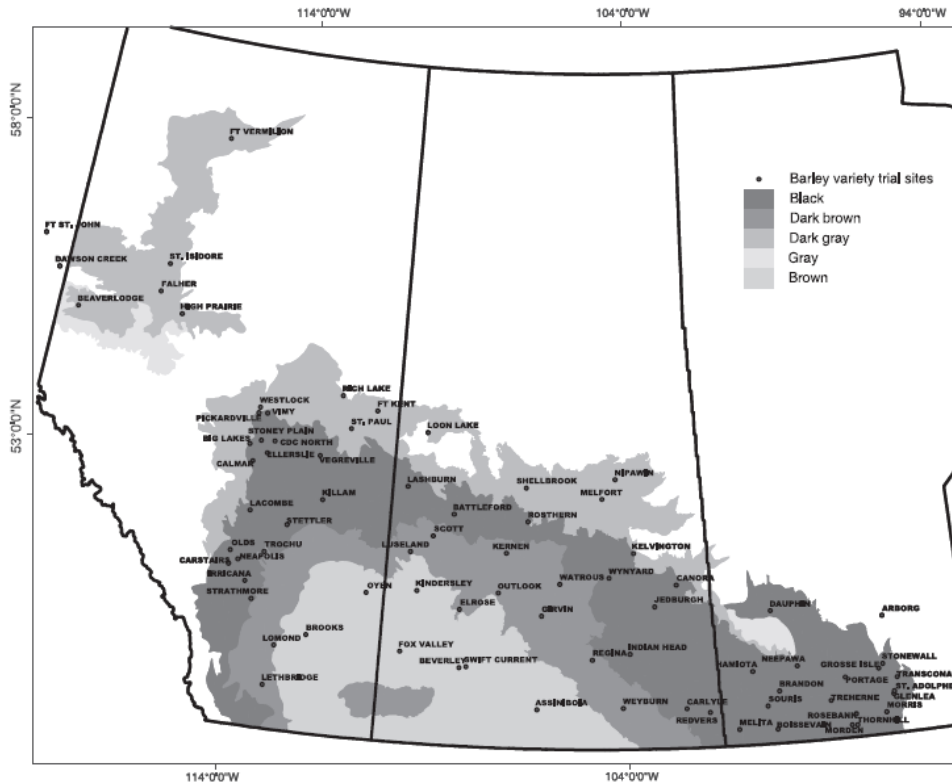
- **Ave stability ( $b=1$ )**
- Low stability ( $b>1$ ): sensitive to environmental changes
- High stability ( $b<1$ ): insensitive to environmental changes



# Isoyield Analysis of Barley Cultivar Trials in the Canadian Prairies

R.-C. Yang, D. Stanton, S. F. Blade, J. Helm, D. Spaner, S. Wright, and D. Domitruk  
J. Agronomy & Crop Science 192, 284—294 (2006)

Isoyield groups not matched  
with soil zones and geography



A modified stability analysis is used to cluster 84 sites  
over the Canadian prairies in 1995 to 2003

# Linear-bilinear approach: seeking for stability patterns

- Based on singular value decomposition (SVD) or two-way PCA
- Different models depending on which of 3 terms (G, E and GxE) SVD is applied
  - AMMI  $\rightarrow$  GxE
  - SREG (GGE)  $\rightarrow$  G + GxE
  - GREG  $\rightarrow$  E + GxE
  - SHMM  $\rightarrow$  G + E + GxE

See Yang et al. (2009, Crop Sci. 49:1564–1576) for review and critiques

# GxE two-way table from 35 soybean trials (Zobel et al. 1988, Agron. J. 80:388-93)

Environment	Genotype							Mean
	EVAN†	WILK	CHIP	HODG	S200	CORS	WELL	
	kg ha <sup>-1</sup>							
A77	2757	2502	2361	2771	2878	3141	2777	2741
C77	2946	2771	1755	2192	1089	1836	1217	1972
V77	1567	1103	2266	2468	2730	2569	2616	2188
V78	1735	1493	1506	2172	2011	2145	1553	1802
A79	3127	2623	2488	3201	3430	2878	2791	2934
C79	2777	2562	1728	1944	1849	1486	1264	1944
G79	2986	2367	2340	3154	2623	3040	2455	2709
R79	1843	1110	1816	2495	2105	1769	2058	1885
V79	1083	578	1278	1500	1964	1661	1715	1397
A80	2919	2784	2582	3208	2703	3100	2219	2788
C80	2596	3248	2280	2710	2172	2260	1432	2385
G80	3901	3194	3376	4096	3887	4250	3517	3746
L80	3706	3820	2993	3732	3739	3161	3215	3481
D80	1937	1580	1580	2374	1997	2609	1890	1995
R80	2199	1870	2199	2966	2461	2327	2186	2315
V80	2334	2018	1802	1964	1601	2165	1782	1952
A81	3033	2609	2636	3013	2831	3611	2959	2956
C81	3053	3053	1849	2522	1654	2186	1708	2289
G81	3322	2892	3208	3840	3383	4028	3235	3415
L81	2972	2710	2636	2972	2569	2818	2777	2779
D81	2529	1997	2582	3268	2112	2529	2320	2477
R81	2038	1385	2347	2798	2616	2966	2757	2415
V81	3026	3127	2387	2367	2361	2461	2260	2570
A82	2186	1870	1883	2441	2441	2562	2239	2232
L82	2663	1957	2535	2798	3241	3147	2737	2725
G82	3652	3295	2724	3712	3901	3921	3322	3504
V82	2273	1910	1123	1755	1184	1399	1345	1570
A83	2582	2125	2018	2313	2192	2058	1970	2180
I83	1278	1029	1688	1701	2105	1964	1856	1660
G83	4499	4015	3329	4620	3564	4062	3625	3959
A84	3161	2717	3188	3860	3376	3423	3544	3324
N84	2993	2603	2630	3221	3369	3201	2703	2960
C84	3181	2690	2448	3060	2576	2784	2670	2773
I84	1950	1701	2078	2260	2145	2246	1917	2042
G84	4015	3329	3329	3961	4277	4015	3692	3803
Mean	2709	2361	2313	2841	2604	2737	2409	2568

-35 environments  
-7 genotypes  
-245 GxE terms



# Problems with Linear regression analysis of GxE

Table 4. Finlay-Wilkinson regression analysis for soybean yields.

Source	df	Sum of squares	Mean square	P
		$\times 10^{-3}$		
Trials	244	139 991	574	5.19***
Genotype	6	8 992	1 499	13.55***
Environment	34	105 558	3 105	28.07***
GE Interaction	204	25 442	125	1.13 NS
Joint Regr.	1	487	487	4.40*
Gen. Regrs.	5	133	27	0.24 NS
Env. Regrs.	33	1 394	42	0.38 NS
Residual	165	23 427	142	1.28*
Error	667	73 775	111	
Total	911	213 766	235	

\*,\*\*\* = P < 0.05 and 0.001, respectively.

Zobel et al. 1988  
(Agron. J. 80:388-93)

- Regression SS explains only 7.9% of the total GxE variation, but with 19% df
  - $(487+133+1394)/25442 = 0.079$
  - $(1+5+33)/204 = 0.19$

**So the linear regression model is parsimonious, but not effective in capturing the total GxE variation**

# AMMI model for GxE

Table 5. AMMI analysis of variance for soybean yields.

Source	df	Sum of squares	Mean square	P
		$\times 10^{-3}$		
Trials	244	139 991	574	5.19***
Genotype	6	8 992	1 499	13.55***
Environment	34	105 558	3 105	28.07***
GE Interaction	204	25 442	125	1.13 NS
PCA 1	39	18 075	463	4.19***
Residual	165	7 367	45	0.41 NS
Error	667	73 775	111	
Total	911	213 766	235	

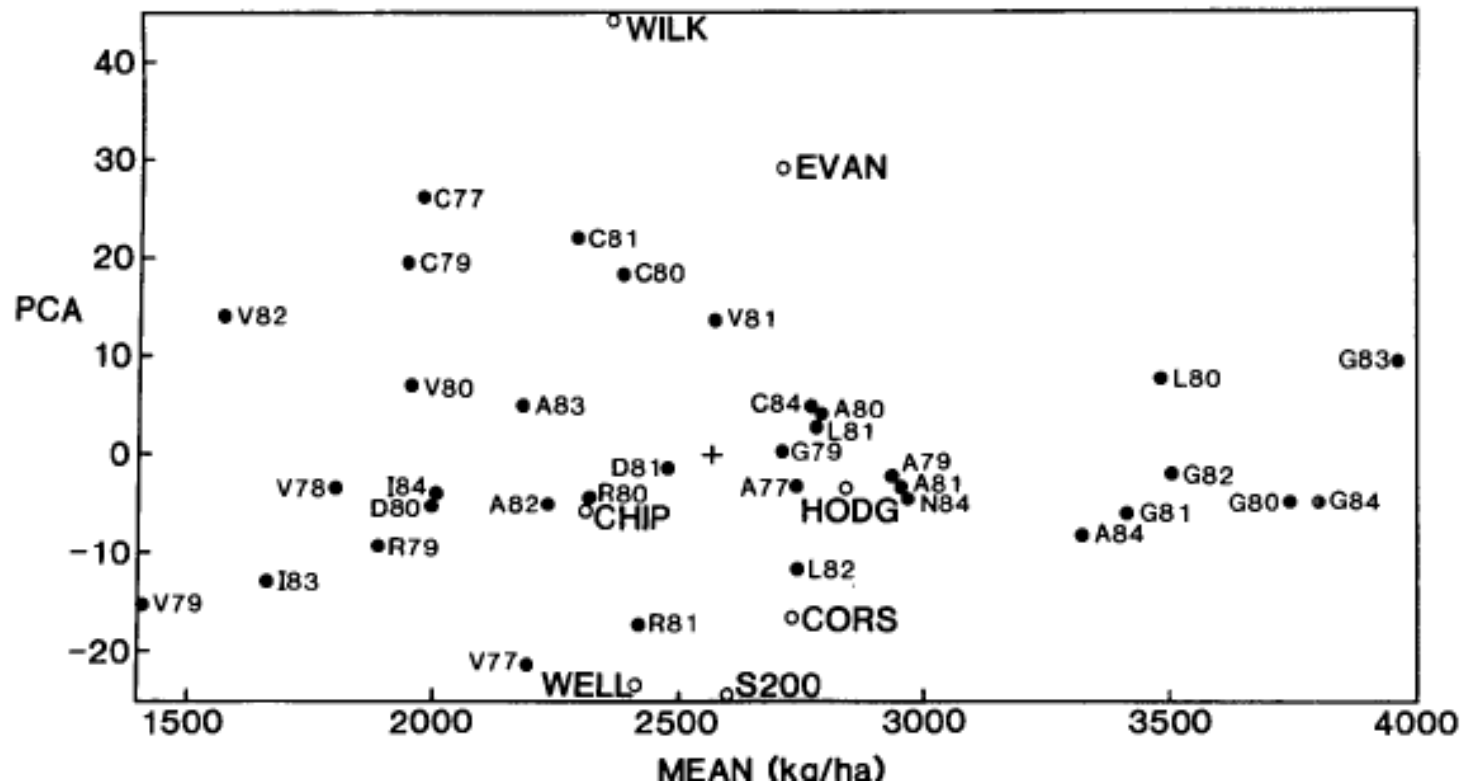
\*\*\* =  $P < 0.001$ .

- AMMI SS explains ~70% of the total GxE variation, but with only 19% df
  - $18075/25442 = 0.68$
  - $39/204 = 0.19$

**So the AMMI model is both parsimonious and effective in capturing the total GxE variation**

Zobel et al. 1988  
(Agron. J. 80:388-93)

# Graphical interpretation of GxE under AMMI model



- Same signs of PC scores for G and E  $\rightarrow$   $G \times E > 0$ ; different signs  $\rightarrow$   $G \times E < 0$ .
- Small  $G \times E$  if PC scores for G and E are close to zero.

# Nonlinear approach: Capturing more GxE variation

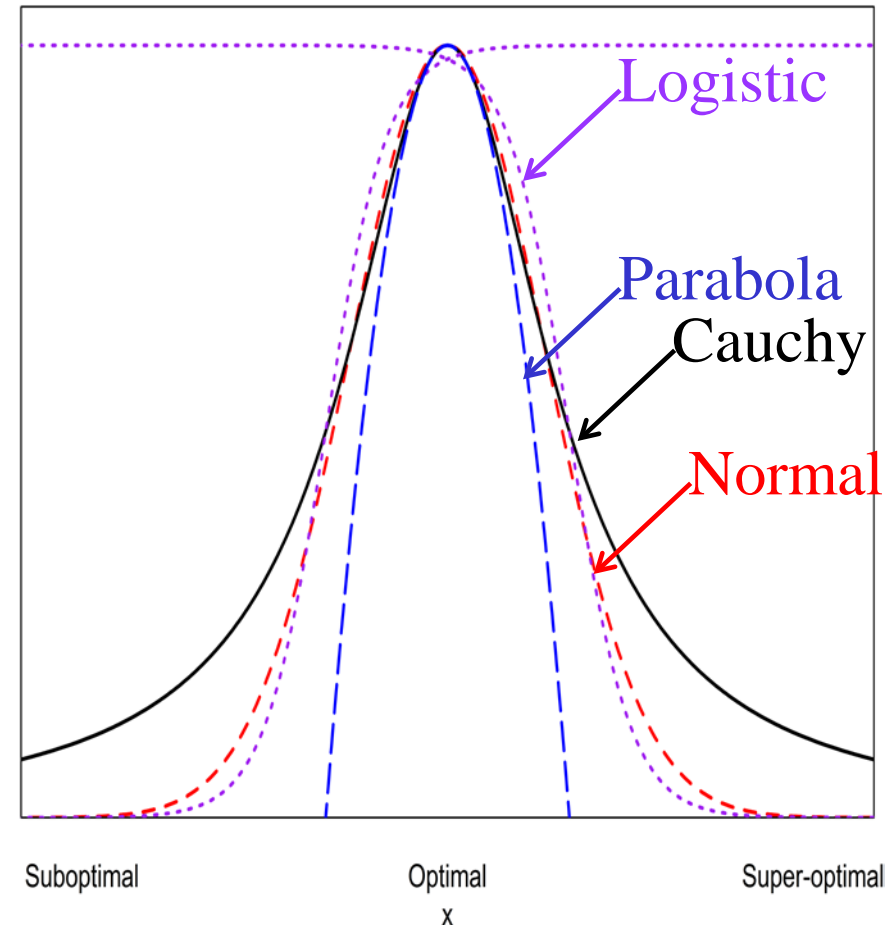
- Some well known nonlinear functions
  - **Quadratic (parabola) function** to model (i) yield response to field water availability (McKenzie et al. 2004), and (ii) response to climate (Rehfeldt et al. 1999)
  - **Cauchy function** to model (i) yield response to planting density (Baker 1988), and (ii) response to geographical gradient (Raymond and Lindgren 1990; Lindgren and Ying 2000)
  - **Logistic function** to model the relationship of plant growth with age (West et al. 2001)
  - **Normal (Gaussian) function** to model response to environmental gradient (Roberds and Namkoong 1989).



# Analysis of linear and non-linear genotype $\times$ environment interaction

Rong-Cai Yang<sup>1,2\*</sup>

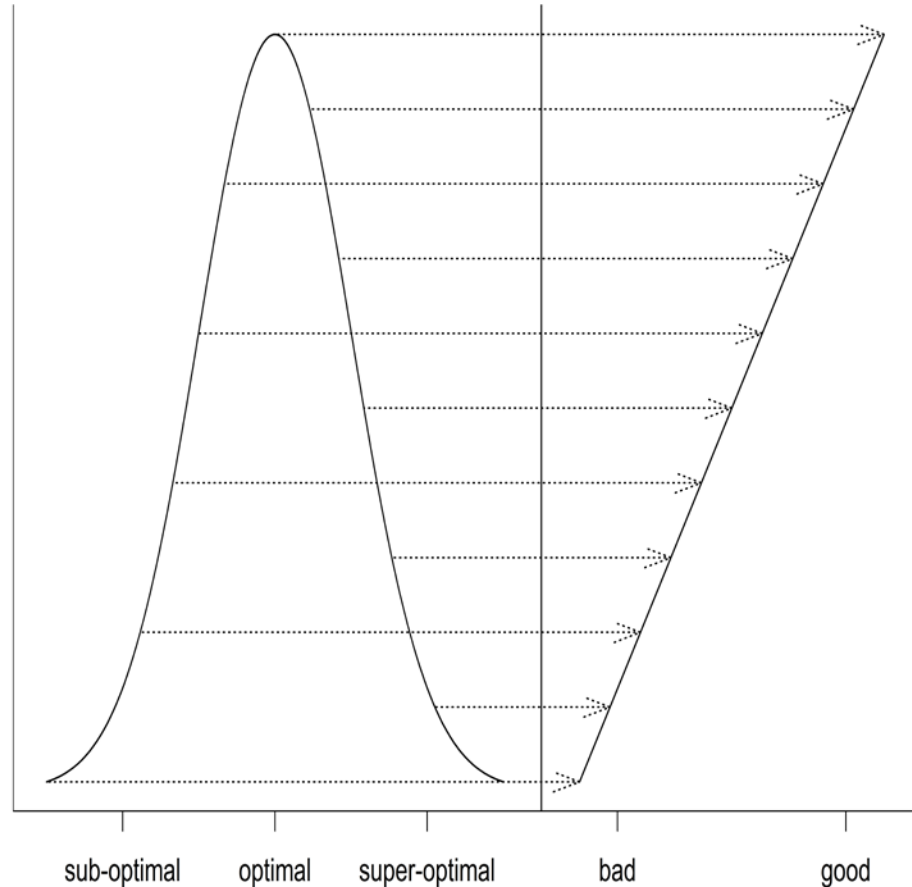
- Responses near the optimum are indistinguishably similar
- Differ only when environment is not good (suboptimal) or too good (super-optimal)
- Efficiency (suboptimal) vs. tolerance (super-optimal)



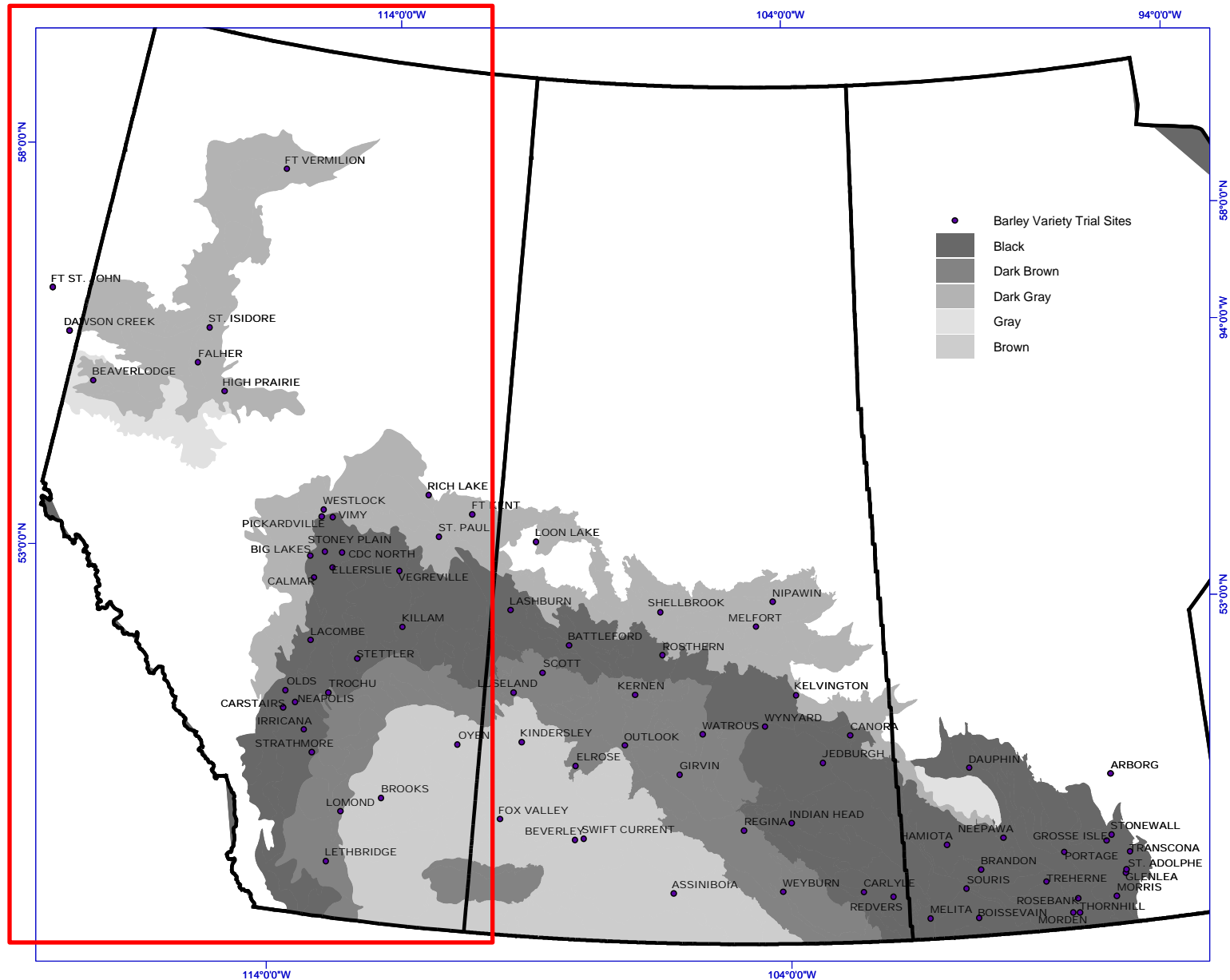
Modified from Lindgren and Ying (2000)

# Linear vs. nonlinear functions

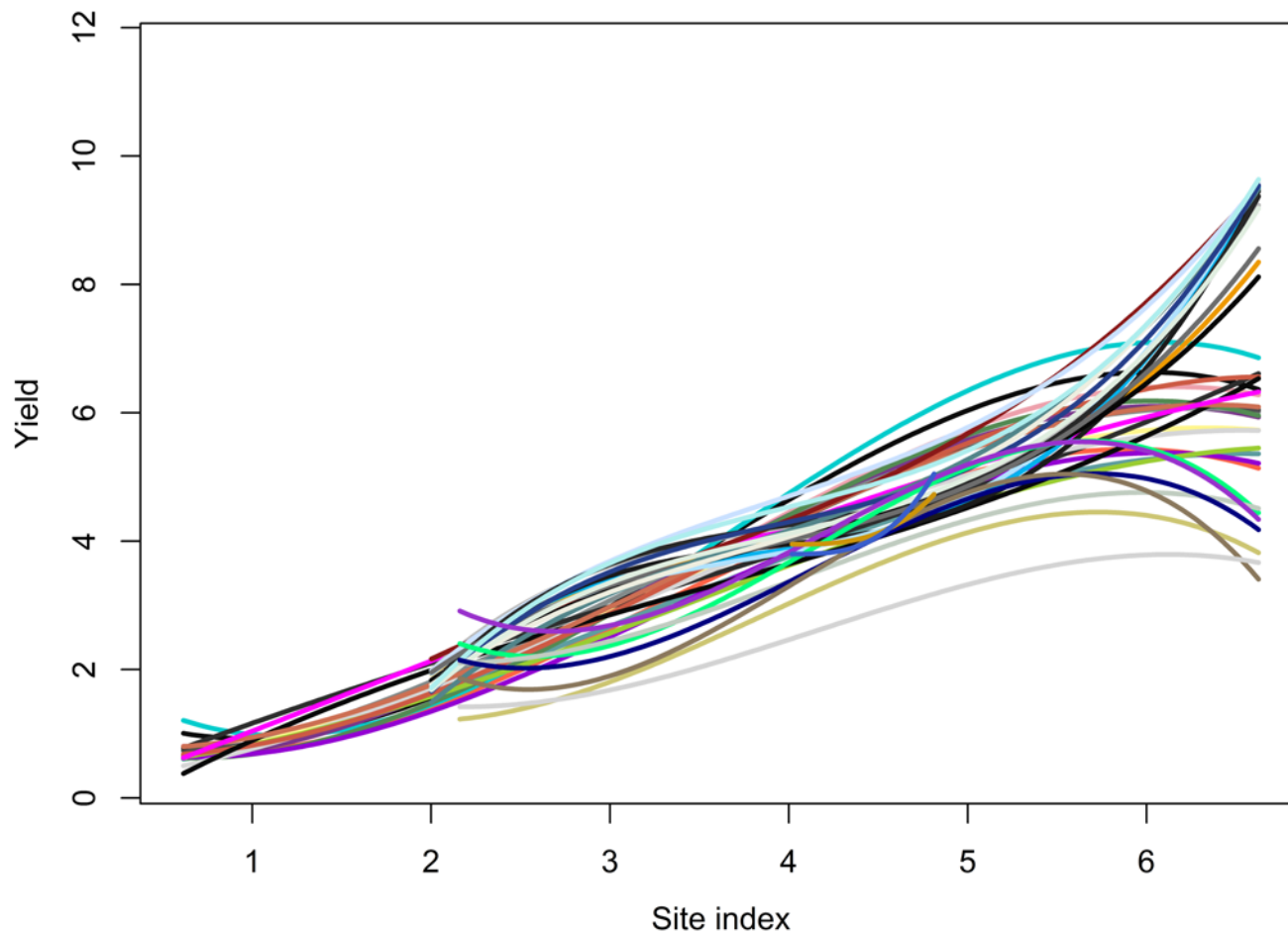
- Linear model can't tell difference between suboptimal and super-optimal
- Linear fit to a nonlinear function would 'force' a reduction in environmental variation (range)



# Distribution of barley variety testing sites across the Canadian prairies (1995-2003)



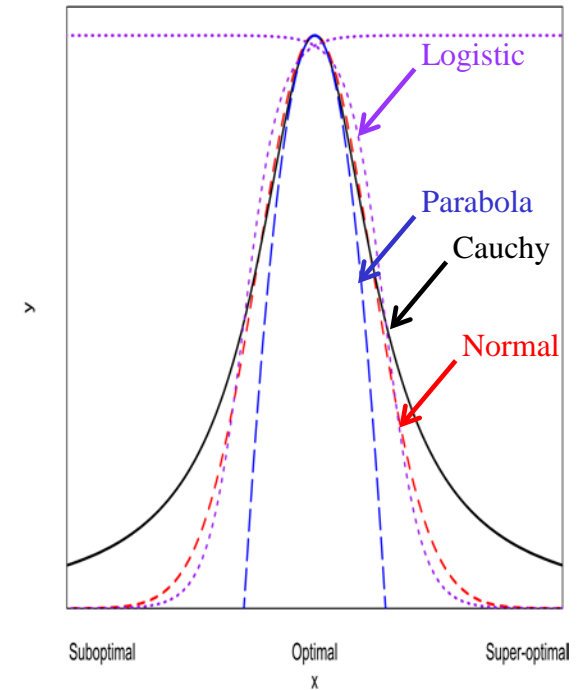
## 2002 Barley variety trials in Alberta: 42 varieties tested at each of 16 sites





# % of total GxE variation for barley cultivar trials in Alberta

	Linear	Logistic	Parabola	Normal	Cauchy
1995	8.49	7.52	11.10	11.47	20.17
1996	8.84	7.32	14.14	13.06	25.28
1997	6.72	5.88	11.81	9.97	12.54
1998	8.40	7.70	13.15	15.12	26.54
1999	14.70	15.75	20.41	20.85	36.56
2000	5.91	8.34	8.67	14.30	32.39
2001	6.95	11.77	13.16	35.04	86.45
2002	23.60	13.17	40.08	33.46	84.87
2003	17.71	14.06	22.51	18.88	37.69
Ave	11.26	10.17	17.23	19.13	40.28



**Cauchy function captures more GxE variation than other functions**

# Genome-wide prediction of environment-specific marker effects

- A huge and complex literature on genome-wide prediction and its application to animal/plant breeding, but all traced back Meuwissen et al (2001) landmark paper
- Key idea: finding weights for all markers when number of markers  $\gg$  number of scored (phenotyped) individuals (“ $p \gg n$ ”).
- A lot of different approaches to do this including rrBLUP, LASSO, BayesA/B, ..., but all involve in some kind of strategies to shrink marker effects towards zero

Copyright © 2001 by the Genetics Society of America

## Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps

T. H. E. Meuwissen,<sup>\*</sup> B. J. Hayes<sup>†</sup> and M. E. Goddard<sup>†,‡</sup>

*<sup>\*</sup>Research Institute of Animal Science and Health, 8200 AB Lelystad, The Netherlands, <sup>†</sup>Victorian Institute of Animal Science, Attwood 3049, Victoria, Australia and <sup>‡</sup>Institute of Land and Food Resources, University of Melbourne, Parkville 3052, Victoria, Australia*

Manuscript received August 17, 2000  
Accepted for publication January 17, 2001



# Analysis of linear and non-linear genotype $\times$ environment interaction

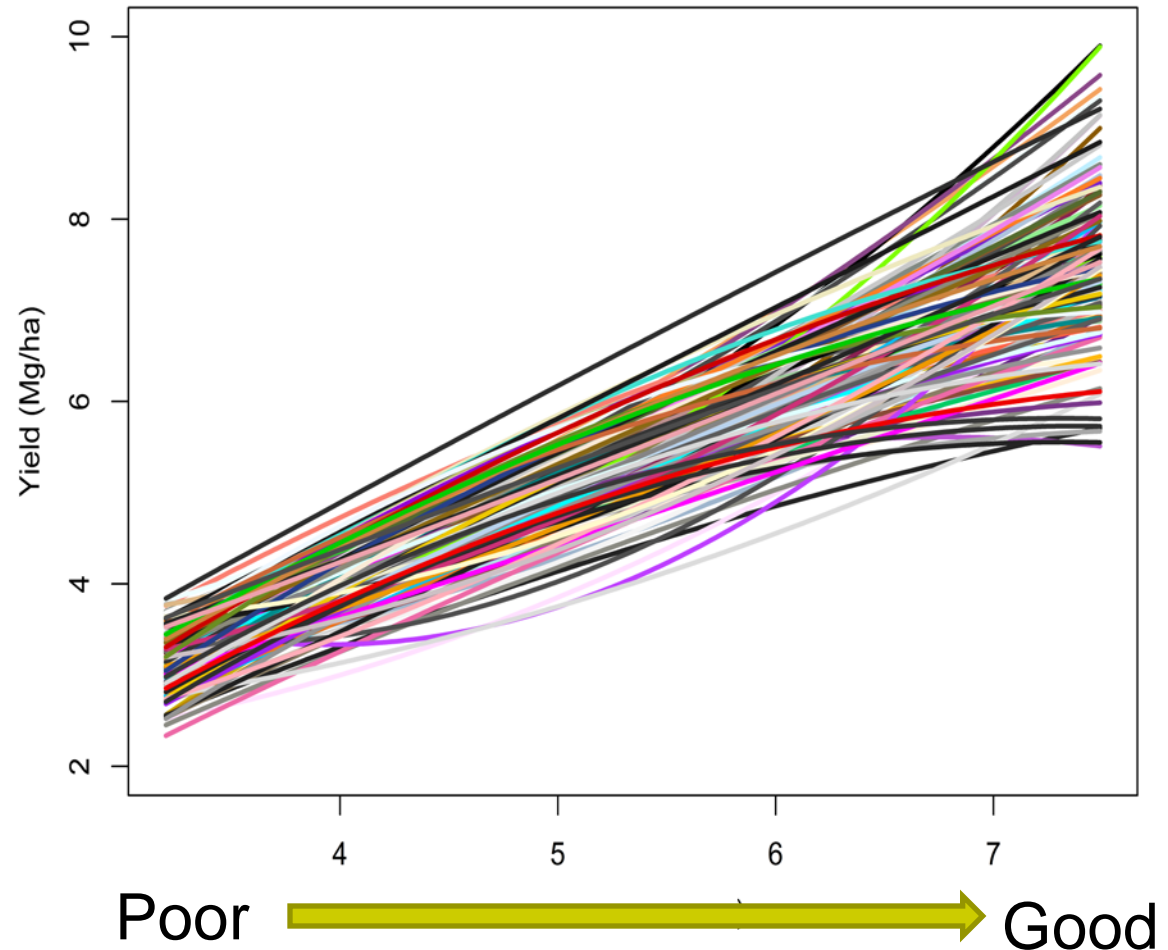
Rong-Cai Yang<sup>1,2\*</sup>

## An example of genomic approach

- A publicly available data set: 150 DH (doubled-haploid) lines developed from a cross between two malting barley varieties (Steptoe  $\times$  Morex) for the North American Barley Genome Mapping Project (NABGMP) (<http://wheat.pw.usda.gov>).
- DH lines were tested in 16 environments over N America for yield and 7 other agronomic and malt quality traits.
- 223 RFLP makers mapped over the genome: 37, 37, 31, 33, 29, 22 and 34 makers were mapped on seven chromosomes.

# Linear response of 150 barley DH lines to environmental index

- Responses are more or less linear
- Variation is greater in 'good' environments than in 'poor' environments.



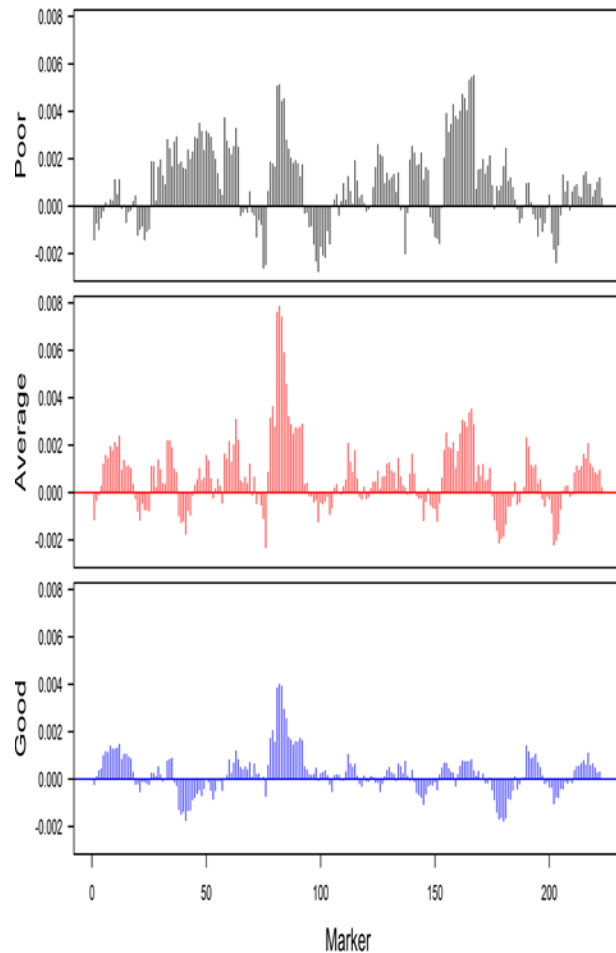
# How to shrink marker/QTL effects?

Shrinkage is needed as  $p(223) > n(150)$  in our case

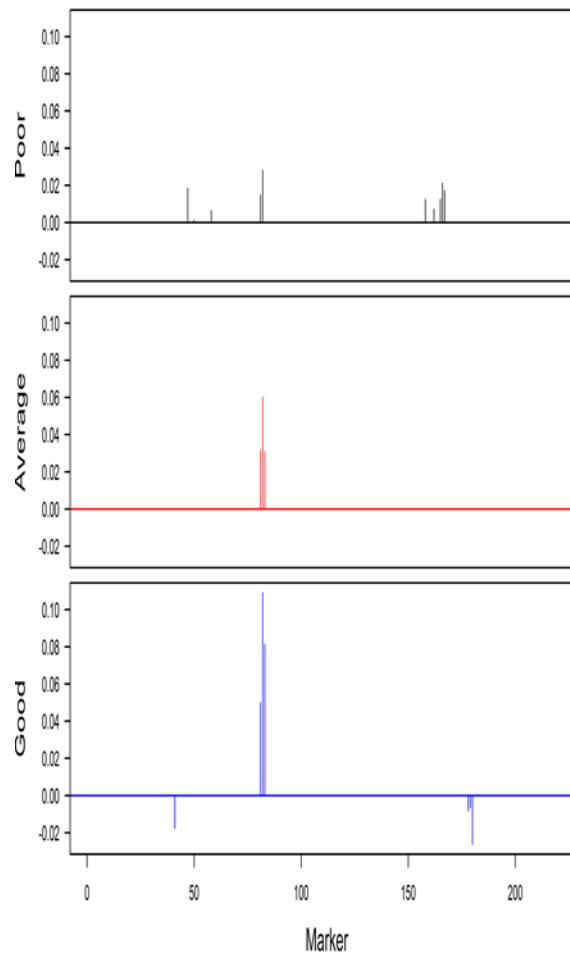
- We used the general regularized regression of Friedman et al. [*J Stat Softw.* 2010 ; 33(1): 1–22] with the **elastic-net penalty** ( $P_\alpha$ ) as a compromise between **rrBLUP** penalty ( $\alpha = 0$ ) and **LASSO** (Least Absolute Shrinkage and Selection Operator) penalty ( $\alpha = 1$ )
- rrBLUP is known to shrink marker effects towards each other, allowing them to borrow strength from each other. In the extreme case of  $p$  markers 'shared' with a single QTL, each marker gets identical weights with  $1/p$  the size of the QTL effect.
  - rrBLUP penalty is ideal for Fisher's infinitesimal model (small but equal genic effects  $\Rightarrow$  a Gaussian prior in Bayesian analysis)
- LASSO is insensitive to highly correlated markers, and it tends to pick one and ignores the rest. In the extreme case above, the lasso problem breaks down.
  - LASSO penalty corresponds to a Laplace prior, which expects many coefficients to be close to zero, and a small subset to be larger and nonzero.
- Implemented in a software, **GLMNET/R**

# Genome-wide estimation of marker/QTL effects: Barley yields in good, average and poor environments

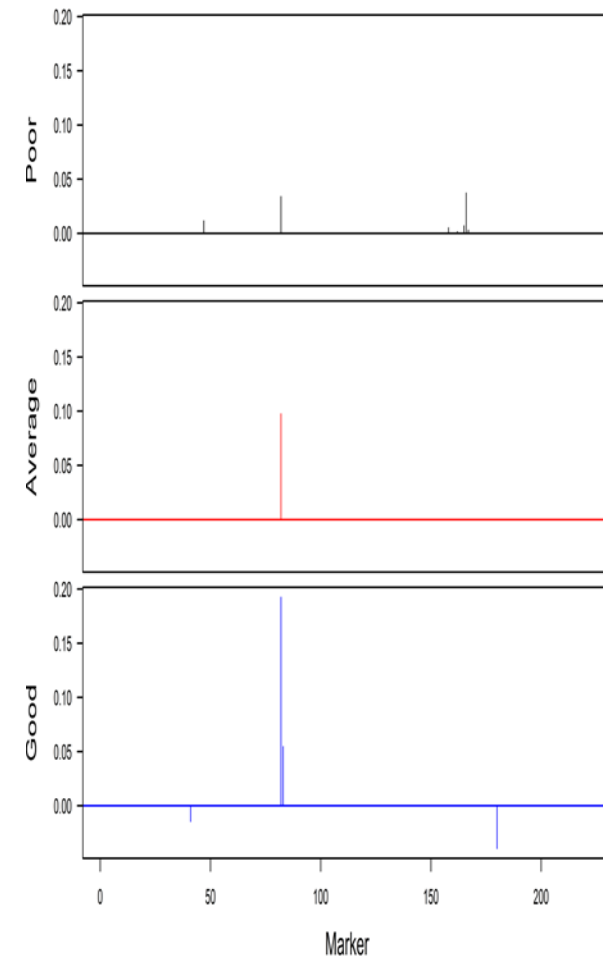
rrBLUP ( $\alpha = 0$ )



Elastic net ( $\alpha = .5$ )



LASSO ( $\alpha = 1$ )

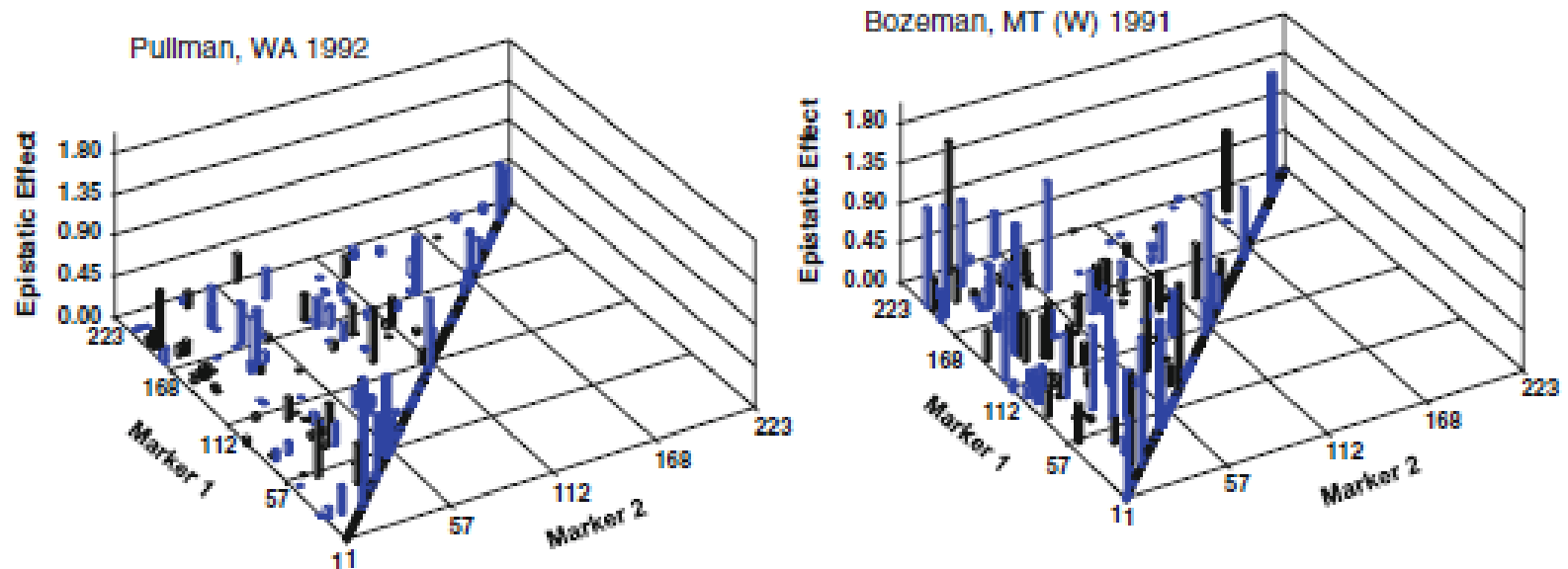


**Environment-specific QTL effects??**

# Epistatic effects vary over environments

Stability of genome-wide QTL effects on malt  $\alpha$ -amylase activity in a barley doubled-haploid population

R.-C. Yang · B. J. Ham

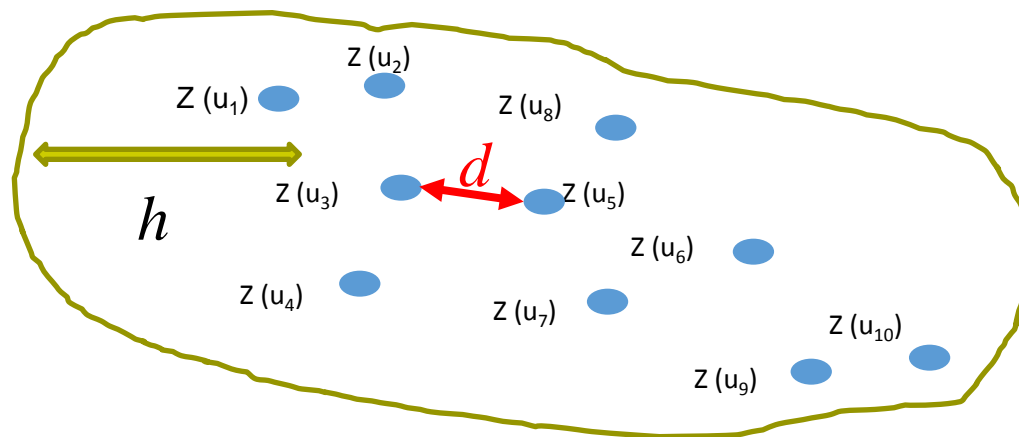


$223(223-1)/2 = 24,753$  pairs of loci for epistatic effects estimated using EBAYES of Xu (2007) in each of nine environments

# Geostatistical approach to GxE

- Make use of location information on individuals to partition the total GxE variability into components due to spatial pattern and residual
- Semivariogram [ $\gamma(h)$ ] = half average of all squared differences between pairs of yield readings at a distance,  $h$ , apart.

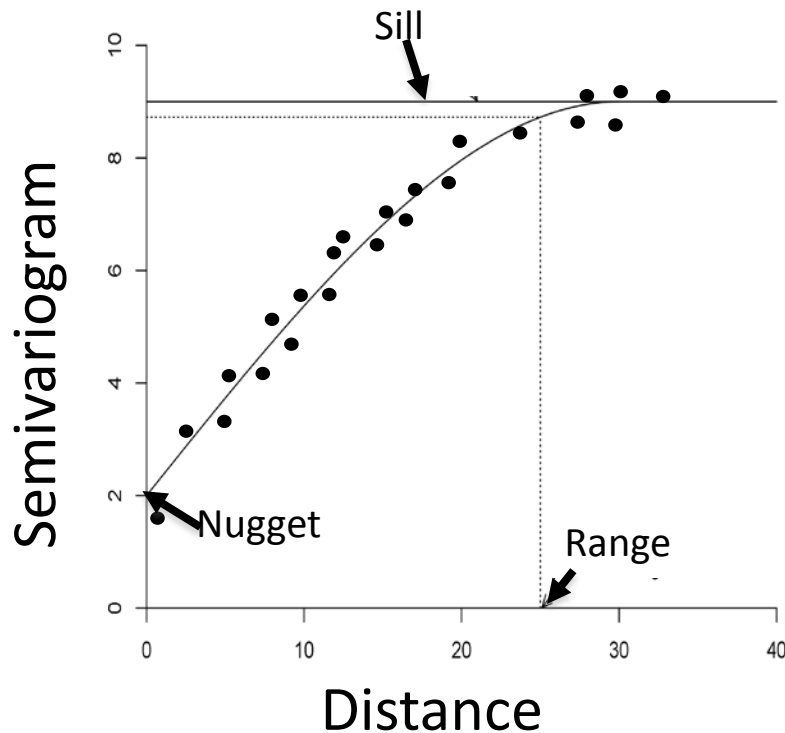
$$\gamma(h) = \frac{1}{2N(h)} \sum_{i \neq j}^{N(h)} [z(u_i) - z(u_j)]^2$$





# Variogram analysis

Determine dependence of GxE variation on geographic distance



Three commonly used models for spatial variability:

- Exponential
- Gaussian
- Spherical

**A similar analysis for determining dependence of GxE variation on difference between climatic or ecological attributes?!!**

# Practical Implications of GxE:

## Performance vs. stability (sensitivity)?

- Performance and stability are both important but for different reasons
  - Increased stability or reduced sensitivity is an important breeding goal in subsistence agriculture because farmers and their families simply cannot afford even a single bad year
  - Improved genotypic stability in forest trees will help mitigate adverse impacts of changing, unpredictable climates in the future
  - Conversely, increased sensitivity is needed for breeding for high-performing genotypes in advanced agronomic management regimes (e.g., “Barley 180” or “Wheat 150” projects in western Canada).

# Take-home messages

- A huge and complex literature on GxE, but the majority has little to do with breeding for predictable and unpredictable environments
- GxE is often 'avoided' by identifying or developing stable genotypes across all environments, but their performance may be compromised at specific environments
- With the advent of genomics and geomatics technologies, it is now possible to exploit GxE by identifying the 'right' genotypes for 'right' environments for maximizing performance

**Thank you!**