

Bayesian Classification and Regression Trees

James Cussens

York Centre for Complex Systems Analysis &

Dept of Computer Science

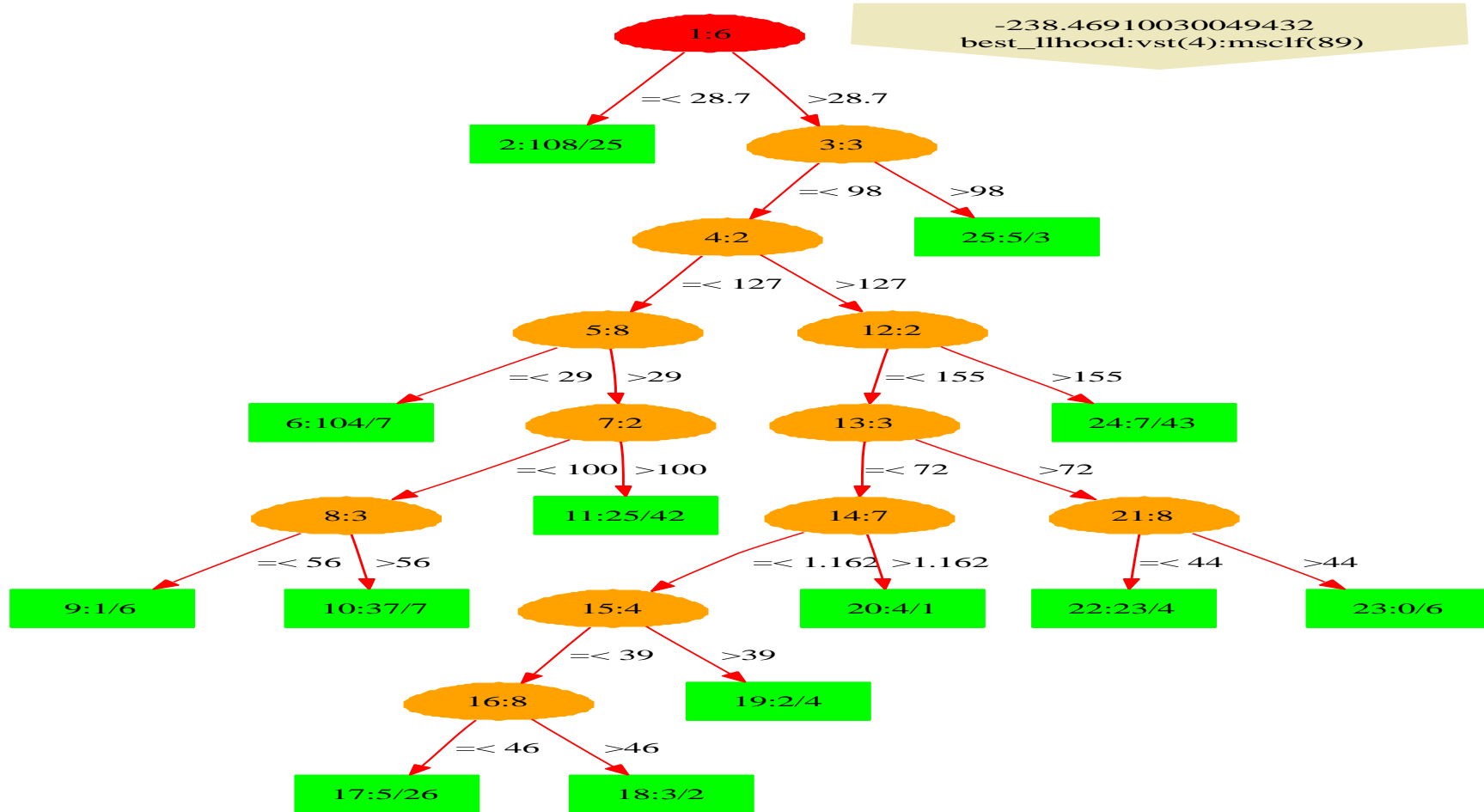
University of York, UK

Outline

- Bayesian C&RT
- Problems for Bayesian C&RT
- Lessons from Bayesian phylogeny
- Results

- **Bayesian C&RT**
- Problems for Bayesian C&RT
- Lessons from Bayesian phylogeny
- Results

Trees are partition models



Bayesian C&RT

Classification trees as probability models

- Tree structure T partitions the attribute space.
- Each partition (= leaf) i has its own class distribution with $\theta_i = (p_{i1}, \dots, p_{iK})$. Let $\Theta = (\theta_1, \theta_2, \dots, \theta_b)$ be the complete parameter vector. for a tree T with b leaves.
- Let x be the vector of attributes for an example, and y its class label.
- (Θ, T) defines a conditional probability model $P(y|\Theta, T, x)$.

The Bayesian approach

- Given
 - Prior distribution $P(\Theta, T) = P(T)P(\Theta|T)$
 - Data (X, Y)
- Compute
 - Posterior distribution $P(\Theta, T|X, Y)$
 - We just care about structure:
$$P(T|X, Y) \propto P(T|X)P(Y|T, X)$$

Defining tree structure priors with a sampler

Instead of specifying a closed-form expression for the tree prior, $P(T|X)$, we specify $P(T|X)$ implicitly by a tree-generating stochastic process. Each realization of such a process can simply be considered a random draw from this prior. (Chipman et al, JASA, 1998)

- Grow by splitting leaves η with a probability $\alpha(1 + d_\eta)^{-\beta}$, where d_η is the depth of η .
- Splitting rules chosen uniformly.

Sampling (approximately) from the posterior

- Produce an *approximate sample* from the posterior $P(T|X, Y)$.
- Generate a Markov chain using the Metropolis-Hastings algorithm.
- If at tree T propose T' with probability $q(T'|T)$ and accept T' with probability $\alpha(T, T')$.

$$\alpha(T, T') = \min \left\{ \frac{P(T'|X, Y) q(T|T')}{P(T|X, Y) q(T'|T)}, 1 \right\}$$

Our proposals

- We propose a new T' by pruning $T^{(i)}$ at a random node and *re-growing according to the prior*, giving:

$$\alpha(T^{(i)}, T') = \min \left\{ \frac{d_{T^{(i)}}}{d_{T'}} \frac{P(Y|T', X)}{P(Y|T^{(i)}, X)}, 1 \right\}$$

where d_t is the depth of T .

- So big ‘jumps’ are possible.

Sometimes it's easy

Kyphosis dataset (81 datapoints, 3 attributes, 2 classes)
50000 MCMC iterations, no tempering:

Tree	$\hat{p}_{\text{seed1}}(T_i)$	$\hat{p}_{\text{seed2}}(T_i)$	$\hat{p}_{\text{seed3}}(T_i)$
T_1	0.08326	0.07898	0.08338
T_2	0.05900	0.06154	0.06170
T_3	0.05574	0.05664	0.05610
T_4	0.02466	0.02724	0.02790
T_5	0.02564	0.02674	0.02504
T_6	0.01494	0.01682	0.01530
T_7	0.01390	0.01410	0.01524
T_8	0.01208	0.01324	0.01288
T_9	0.01212	0.01284	0.01168

Computing class probabilities for new data

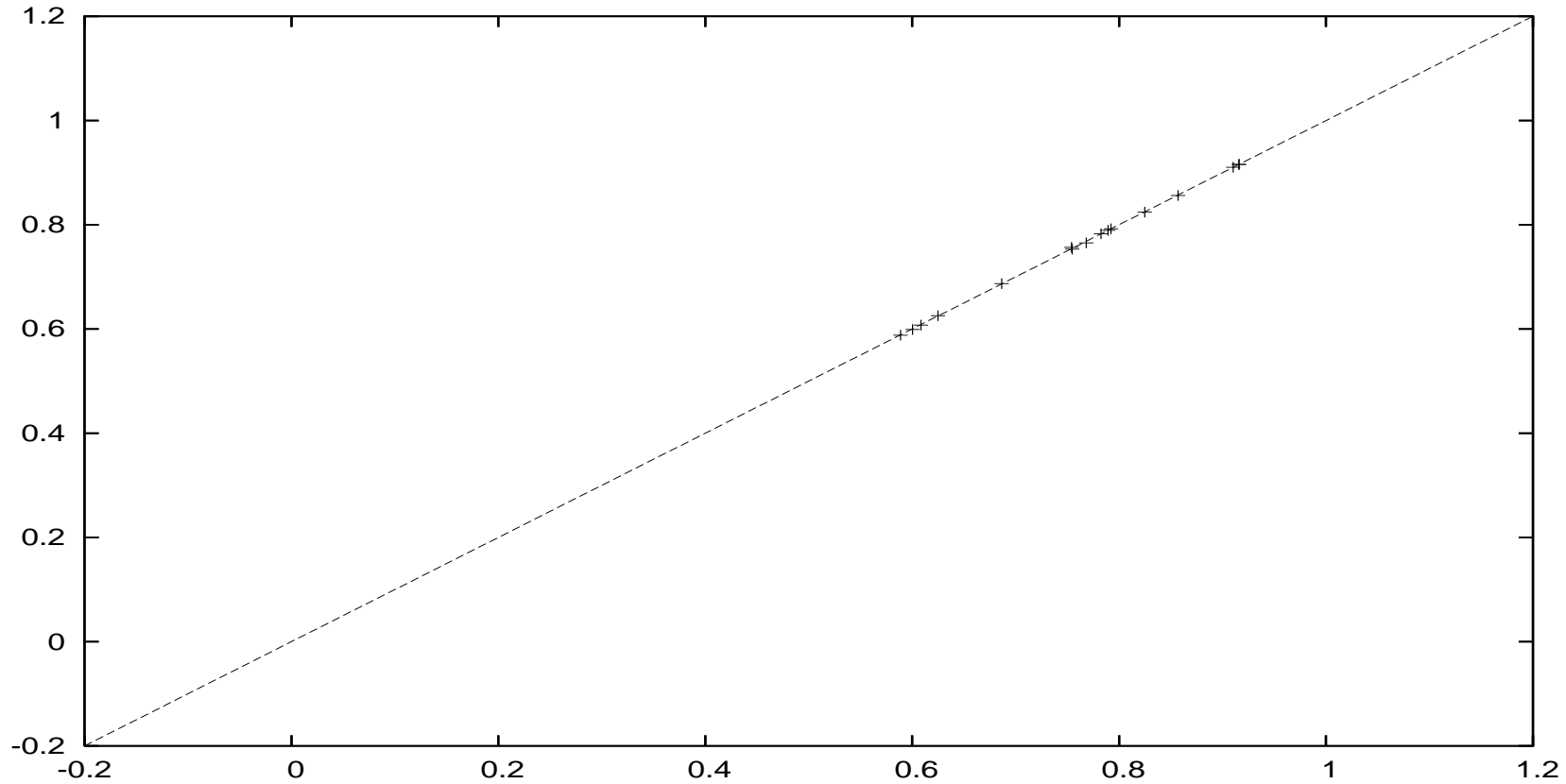
Given training data (X, Y) ,
the posterior probability that x' has class y' is:

$$p(y'|x', X, Y) = \sum_T P(T|X, Y) \int p(y'|x', \Theta, T) P(\Theta|T, X, Y) d\Theta$$

We use the MCMC sample to estimate $P(T|X, Y)$, the rest is analytically soluble.

Comparing class probabilities in an easy case

(tr_uc_rm_idsd_a0_95b1_i50K__s) 512 vs. 883



Dataset=K, iterations=50,000, tempering=FALSE

Bayesian C&RT

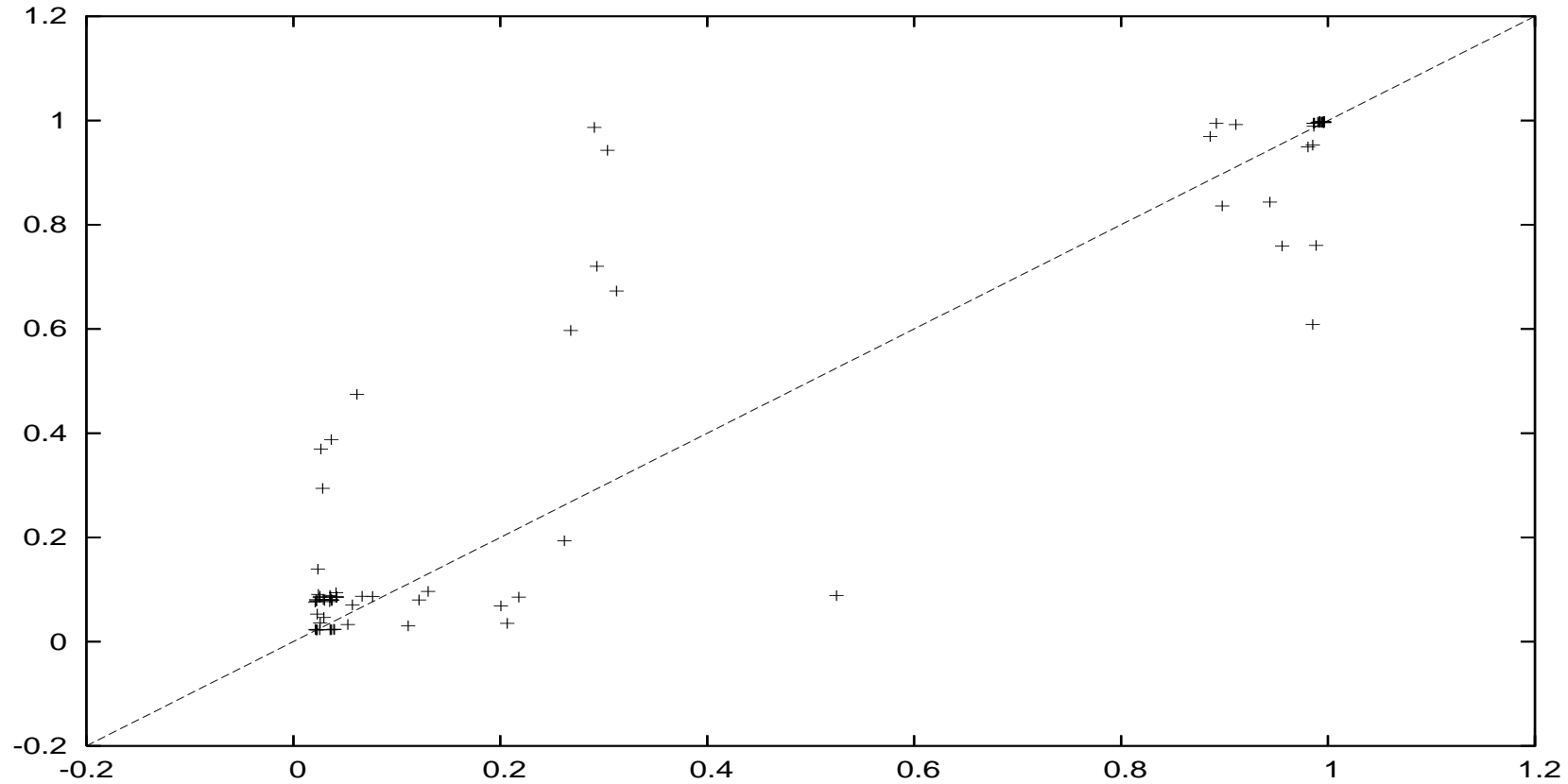
- Bayesian C&RT
- **Problems for Bayesian C&RT**
- Lessons from Bayesian phylogeny
- Results

Usually it's not easy

... the algorithm gravitates quickly towards [regions of large posterior probability] and then stabilizes, moving locally in that region for a long time. Evidently, this is a consequence of a proposal distribution that makes local moves over a sharply peaked multimodal posterior. Once a tree has reasonable fit, the chain is unlikely to move away from a sharp local mode by small steps. ... Although different move types might be implemented, we believe that any MH algorithm for CART models will have difficulty moving between local modes. (Chipman et al, 1998)

Where there is room for improvement

(tr_uc_rm_idsd_a0_95b1_i250K__s) 447 vs. 938

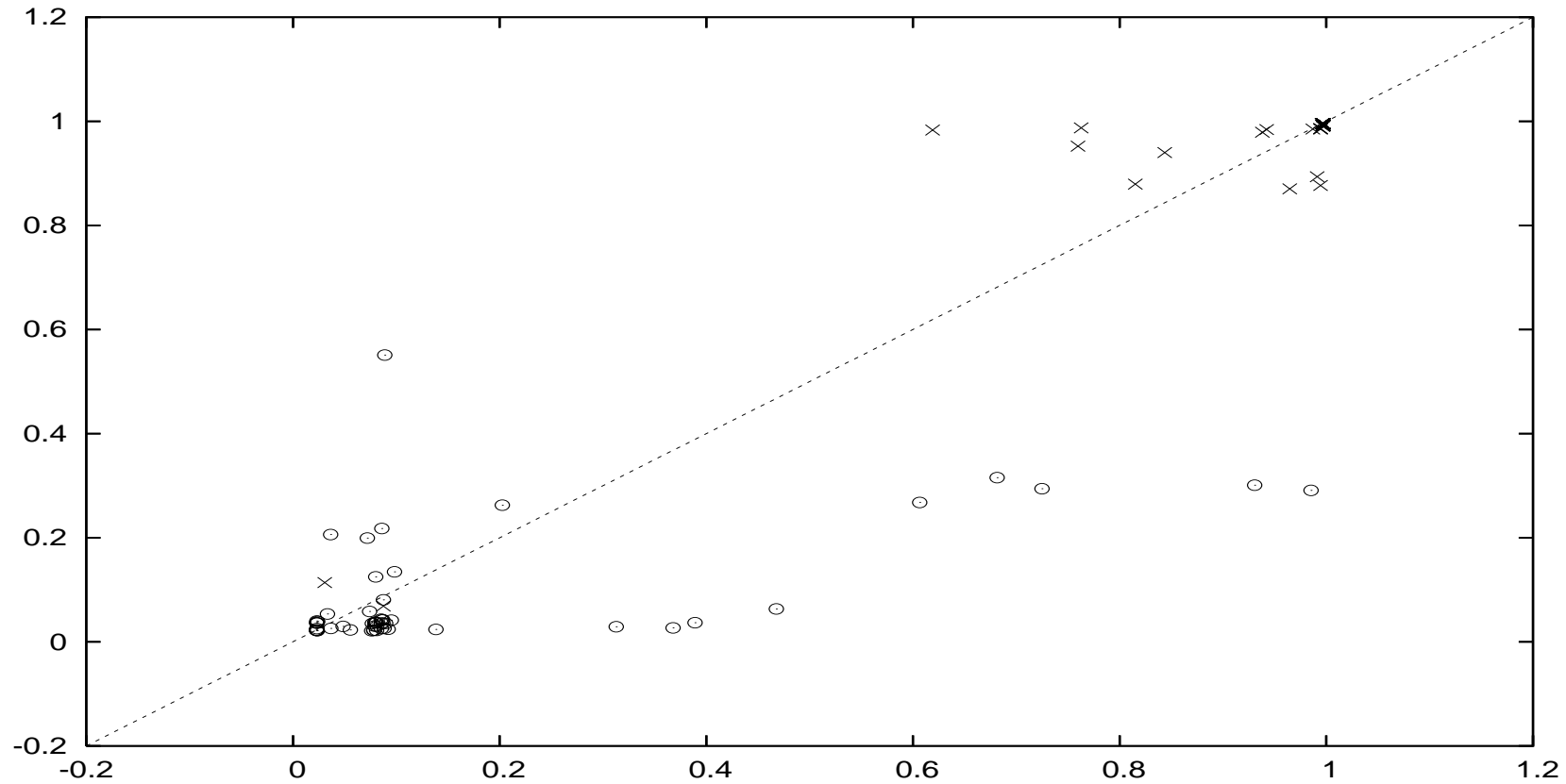


Dataset=BCW, iterations=250,000, tempering=F

Bayesian C&RT

Where there is room for improvement

(tr_uc_rm_idsd_a0_95b1_i250K__s) 938 vs. 447



Dataset=BCW, iterations=250,000, tempering=F

Bayesian C&RT

- Bayesian C&RT
- Problems for Bayesian C&RT
- **Lessons from Bayesian phylogeny**
- Results

The same problem for Bayesian phylogeny

The posterior probability of trees can contain multiple peaks. . . . MCMC can be prone to entrapment in local optima; a Markov chain currently exploring a peak of high probability may experience difficulty crossing valleys to explore other peaks. (Altekar et al, 2004)

MrBayes is at <http://morphbank.ebc.uu.se/mrbayes/>

A solution: (power) tempering

- As well as the ‘cold’ chain with stationary distribution $P(T|X, Y)$,
- Have ‘hot’ chains with stationary distributions $P(T|X, Y)^\beta$ for $0 < \beta < 1$
- And swap states between chains.
- Only states visited by the cold chain count.

Acceptance probabilities for tempering

$$\alpha_{uc}^{\beta}(T^{(i)}, T') = \min \left\{ \frac{d_{T^{(i)}}}{d_{T'}} \left(\frac{P(Y|T', X)}{P(Y|T^{(i)}, X)} \right)^{\beta}, 1 \right\}$$

$$\alpha_{\text{swap}} = \min \left\{ \left(\frac{P(Y|T_2, X)}{P(Y|T_1, X)} \right)^{(\beta_1 - \beta_2)}, 1 \right\}$$

- Bayesian C&RT
- Problems for Bayesian C&RT
- Lessons from Bayesian phylogeny
- **Results**

The small print

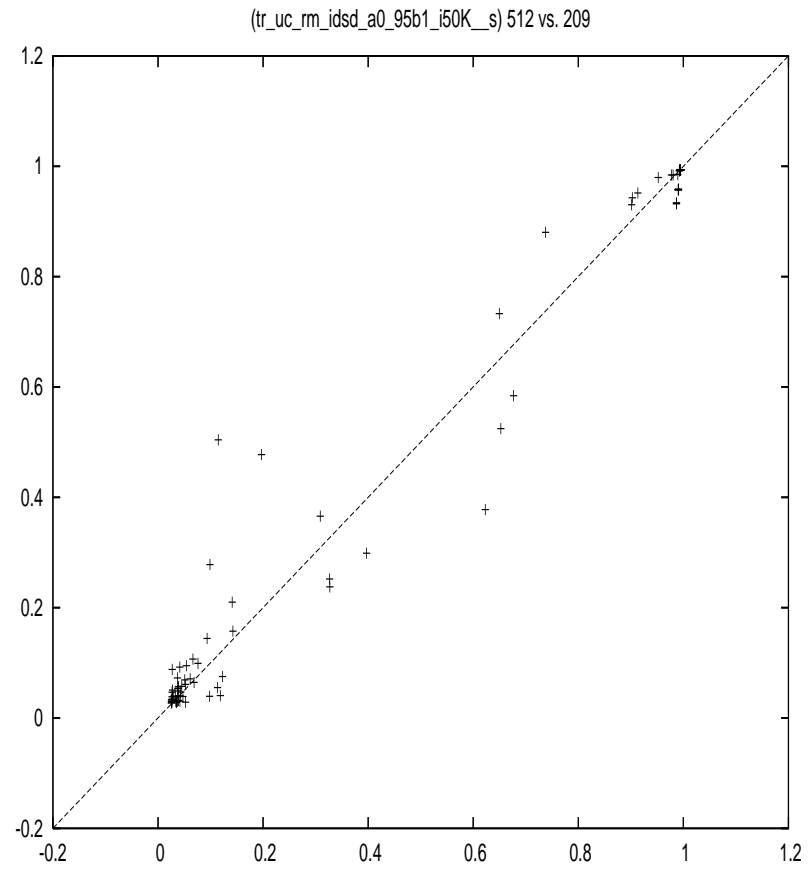
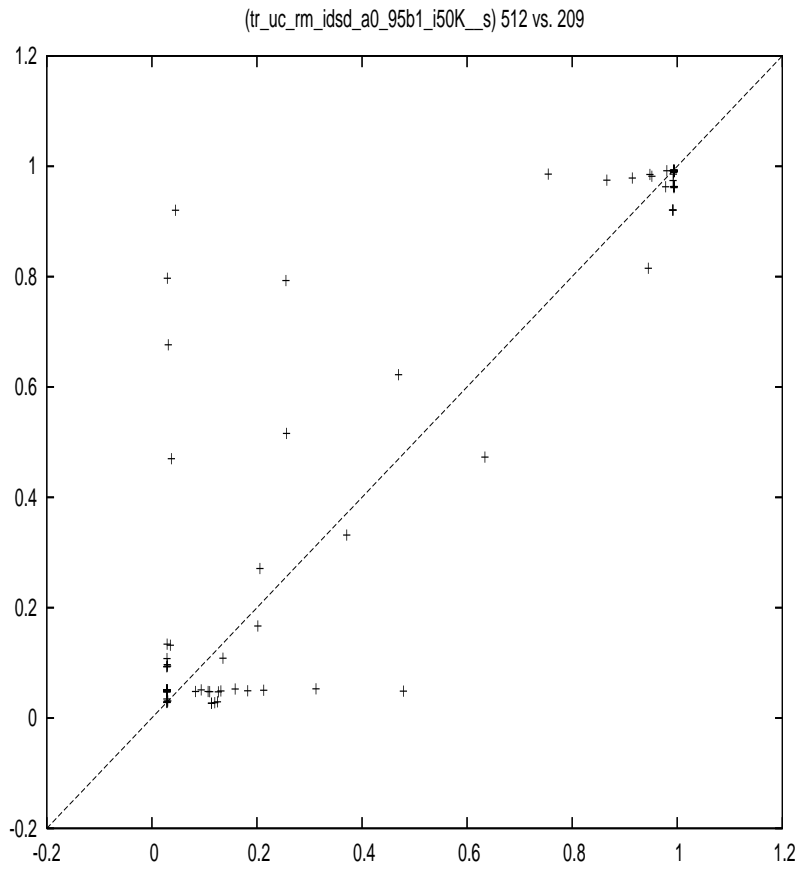
- Copied MrBayes defaults: $\beta_i = 1/(1 + \Delta T(i - 1))$ for $i = 1, 2, 3, 4$, where $\Delta T = 0.2$.

Datasets

Name	Size	$ x $	$ Y $	Pos%(Tr)	Pos%(HO)
K	81	3	2	81.5%	68.8%
BCW	683	9	2	66.2%	60.3%
PIMA	768	8	2	65.4%	64.1%
LR	20000	16	26	3.85%	4.3%
WF	5000	40	3	35.6%	33.4%

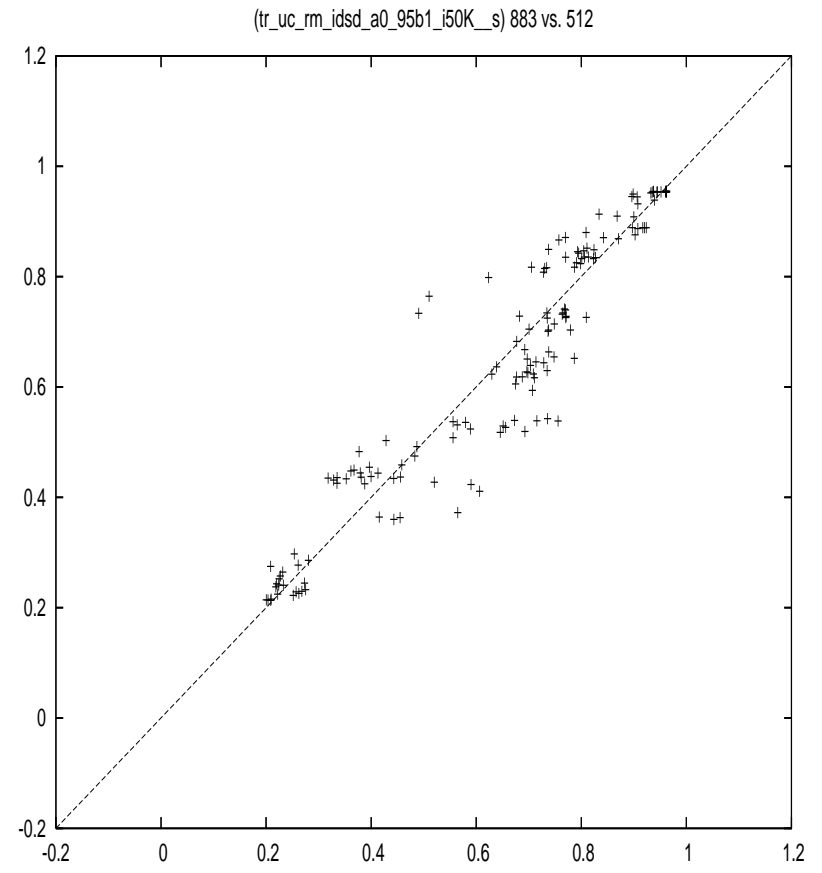
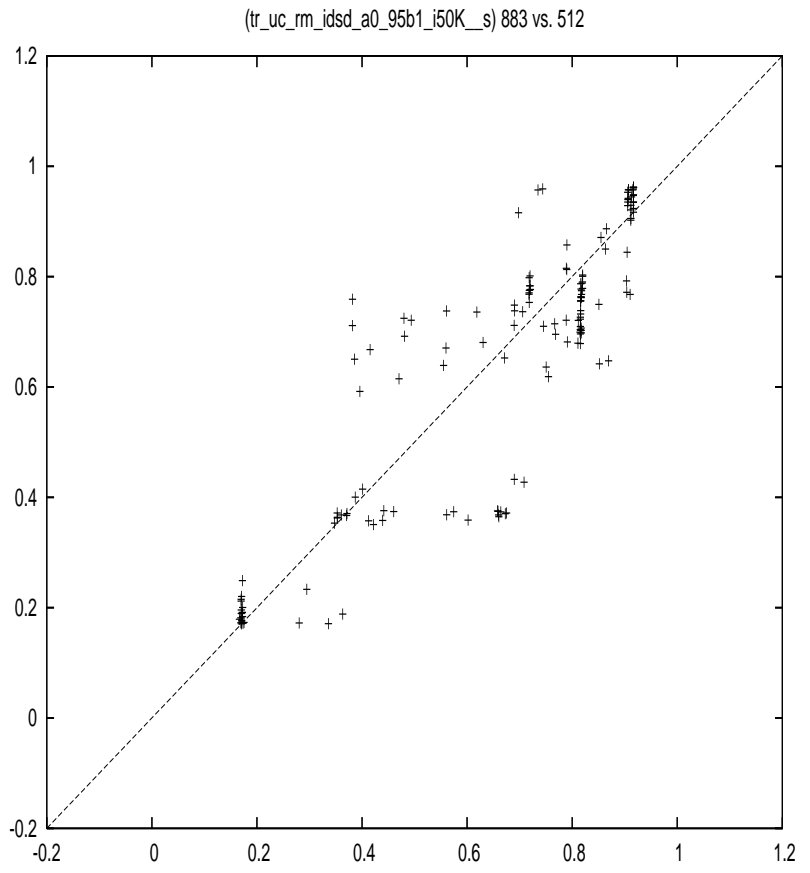
Holdout set (HO) is 20% of the data.

BCW: 50K, Temp=F vs 50K, Temp=T



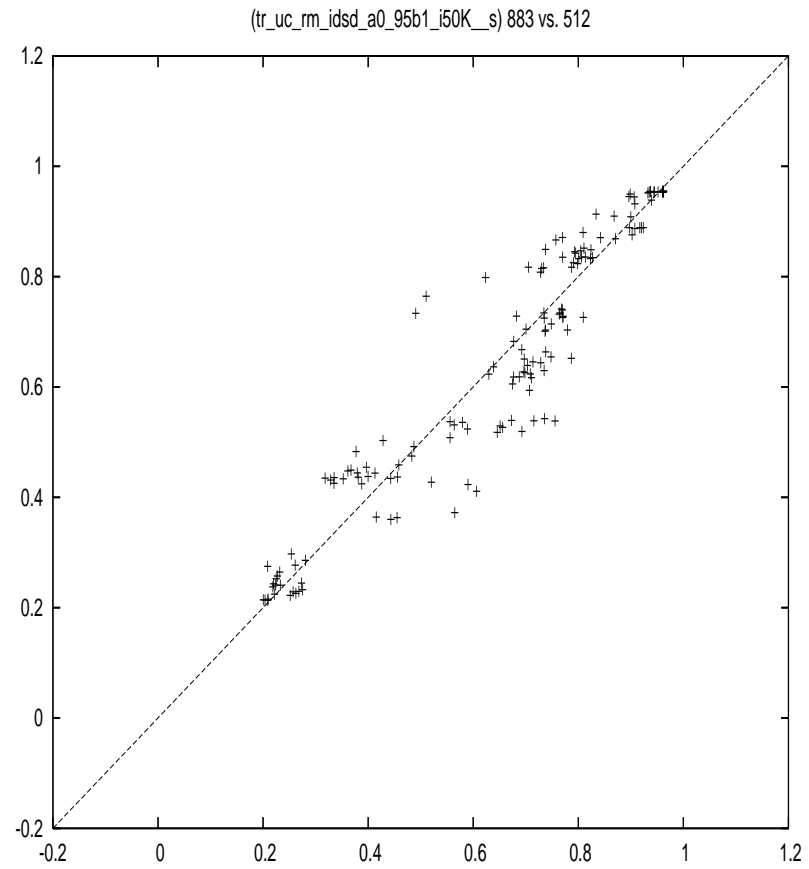
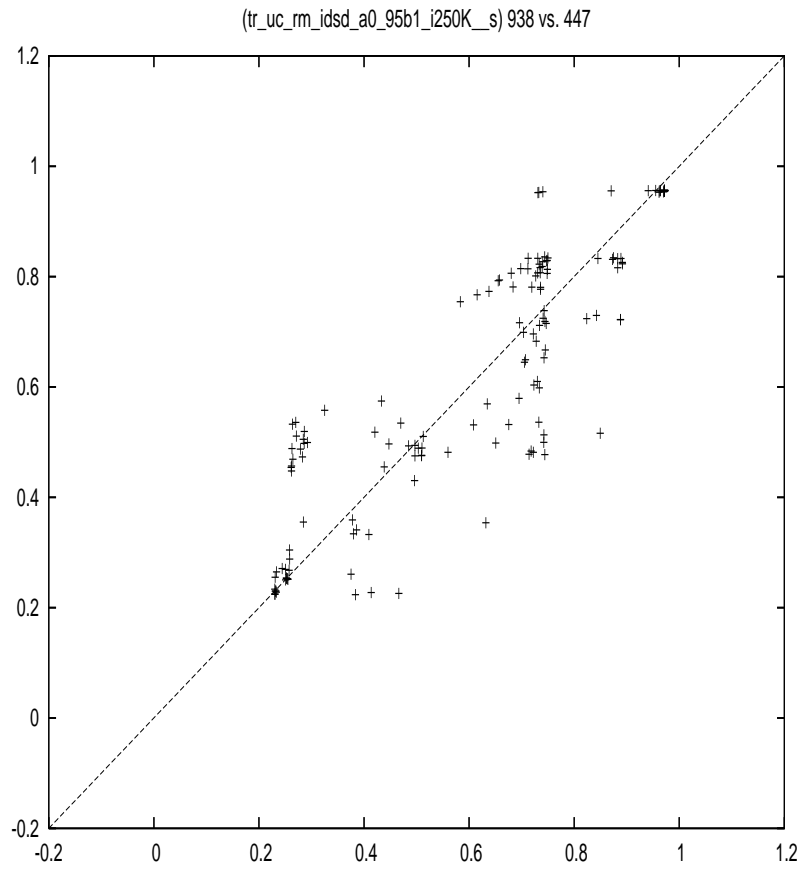
Bayesian C&RT

PIMA: 50K, Temp=F vs 50K, Temp=T



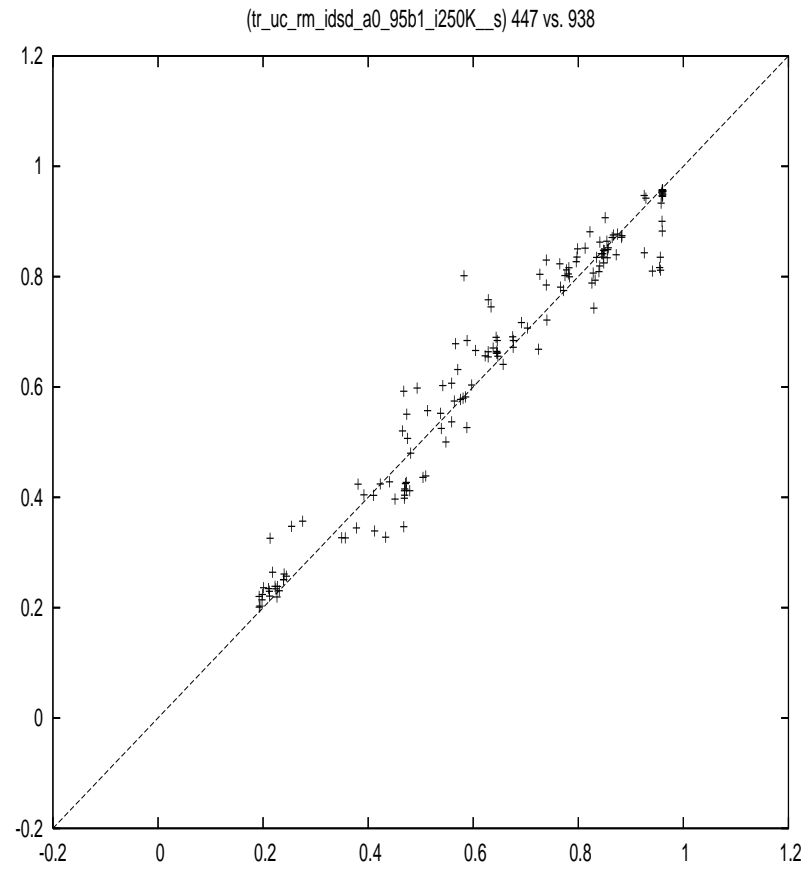
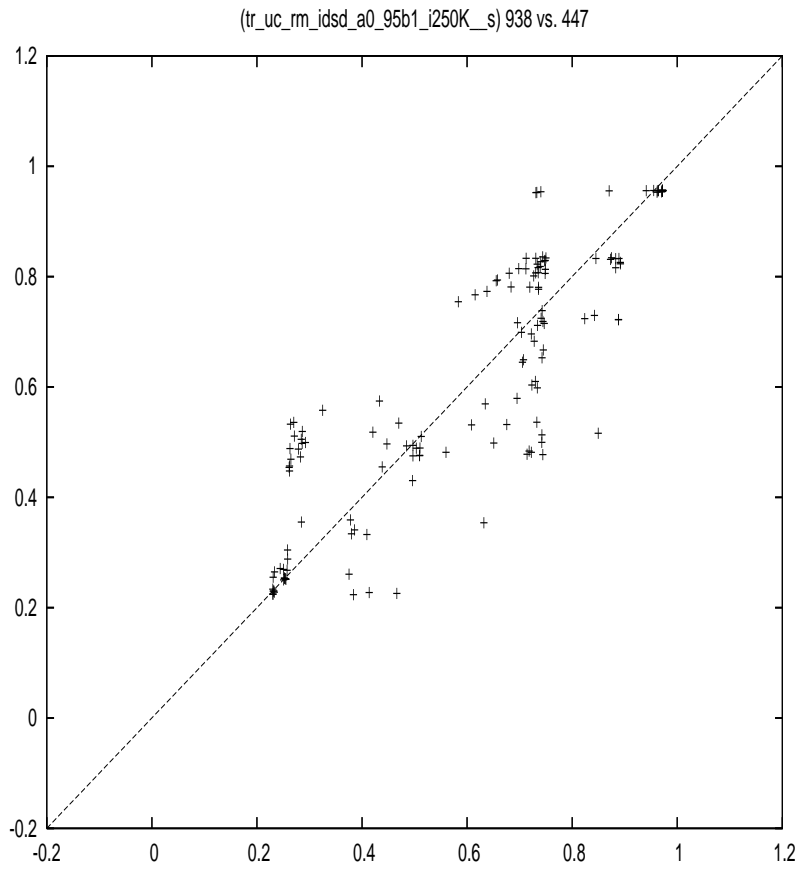
Bayesian C&RT

PIMA: 250K, Temp=F vs 50K, Temp=T



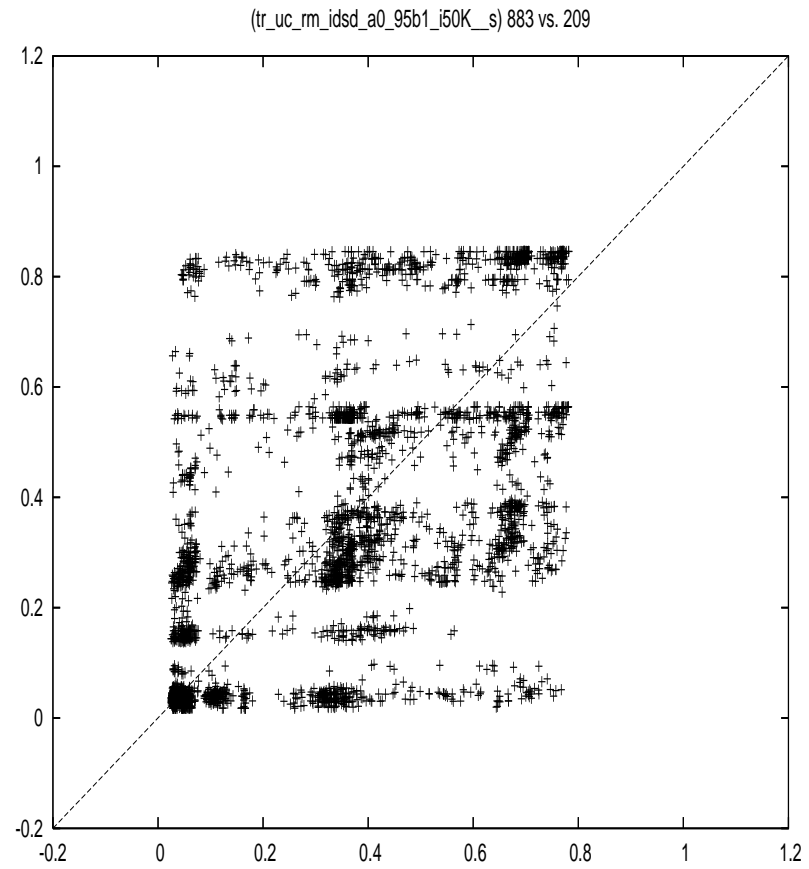
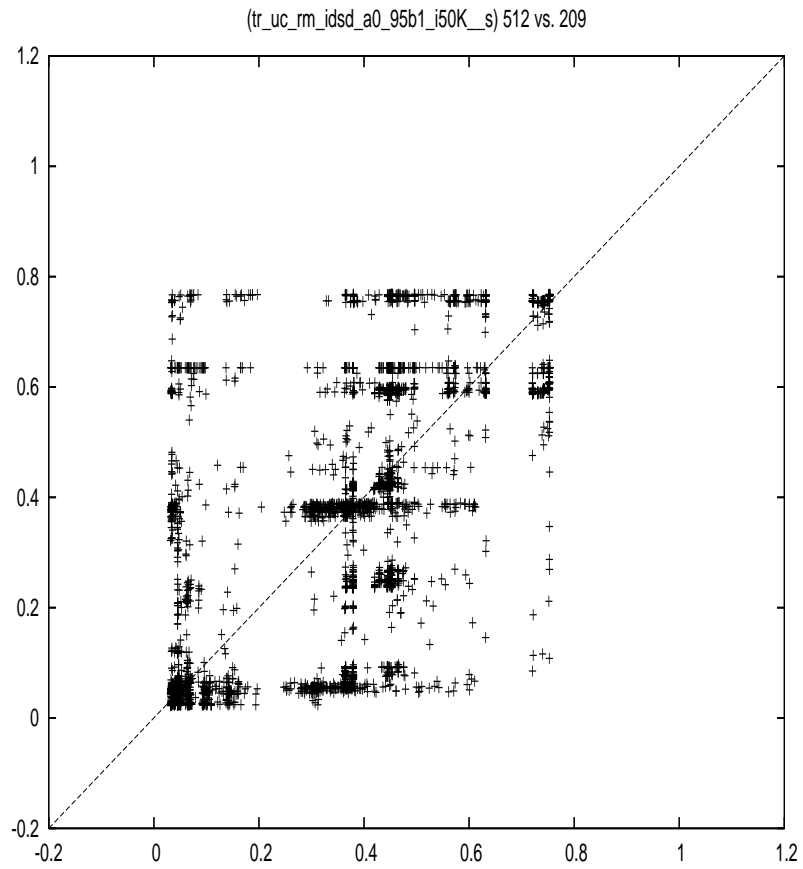
Bayesian C&RT

PIMA: 250K, Temp=F vs 250K, Temp=T



Bayesian C&RT

WF: 50K, Temp=F vs 50K, Temp=T



Bayesian C&RT

Stability of classification accuracy on hold-out set for 3 MCMC runs with and without tempering

Data	Temp=F		Temp=T		rpart	Time per 1000
	$\overline{\text{acc}}$	σ_{acc}	$\overline{\text{acc}}$	σ_{acc}		
K	68.8%	0.0%	68.8%	0.0%	75.0%	5s
BCW	96.1%	1.2%	95.8%	0.3%	95.5%	17s
PIMA	76.9%	3.2%	73.6%	1.6%	76.4%	129s
LR	62.4%	3.6%	66.9%	0.1%	46.1%	2368s
WF	71.0%	3.7%	72.5%	2.9%	74.1%	1151s

Materials

- This SLP and other materials used available from <http://www-users.cs.york.ac.uk/aig/slps/mcmcms/>
- Look in the `pb1/icm105` directory of the MCMCMS distribution.
- Includes scripts for reproducing the figures in this paper.

Future work

- Tempering plus informative priors.
- Currently applying to MCMC for Bayesian nets.

Bayesian Additive Regression Trees

BART uses a sum-of-trees model:

$$Y \sim g_1(x) + g_2(x) + \dots + g_m(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

where each g_i is a regression tree.

- Do MCMC by ‘Gibbs sampling’: get a new tree for g_i conditional on all the others.
- The distribution for the new tree only depends on the residual produced from the other trees.

Getting that posterior

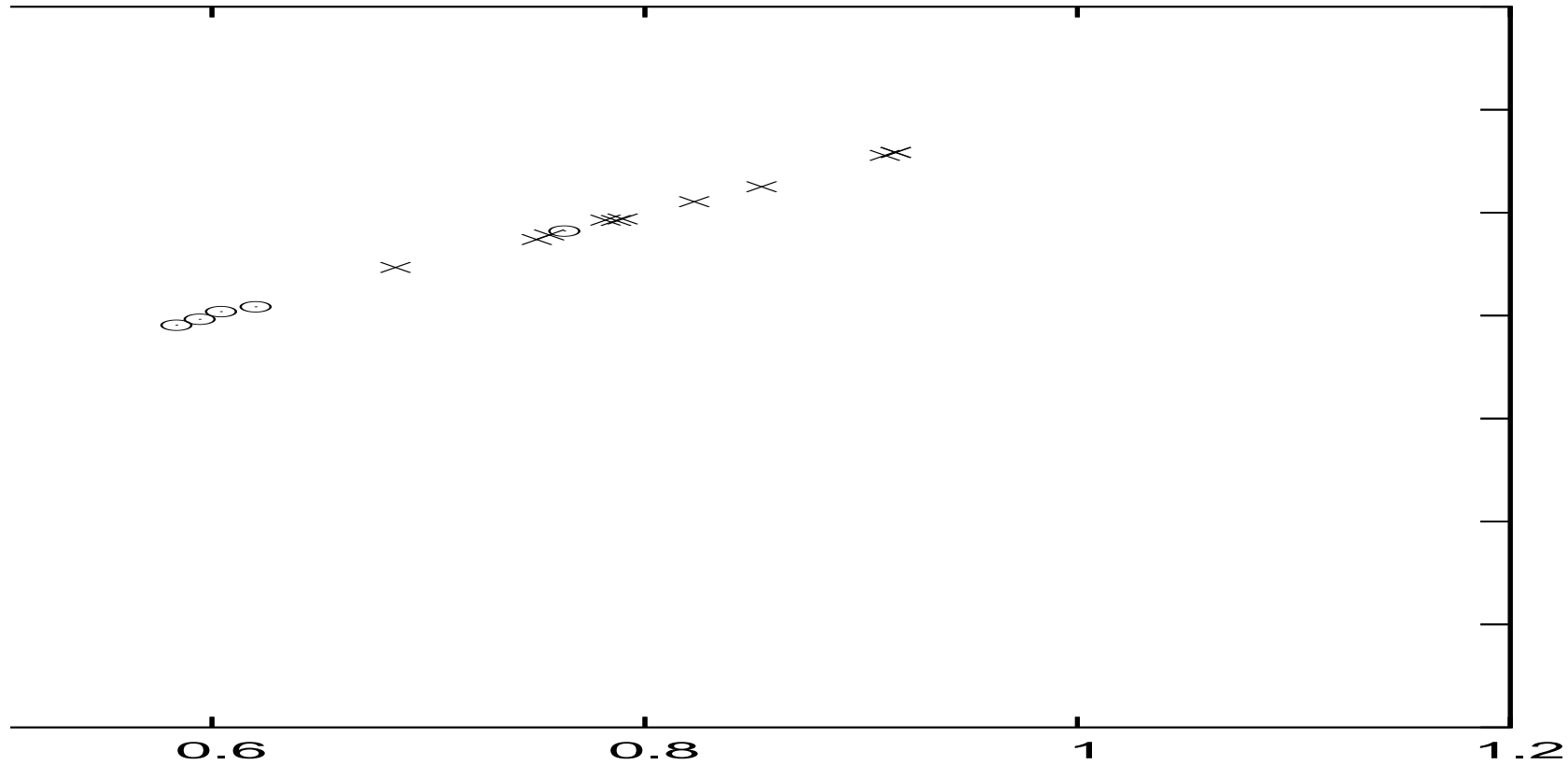
- Bayes theorem: $P(\Theta, T|X, Y) \propto P(\Theta, T|X)P(Y|\Theta, T, X)$
- We restrict attention to tree structures, so integrate away the parameters Θ :

$$\begin{aligned} P(T|X, Y) &= P(T|X) \int_{\Theta} P(Y|\Theta, T, X)P(\Theta|T, X)d\Theta \\ &= P(T|X)P(Y|T, X) \end{aligned}$$

- The *marginal likelihood* $P(Y|T, X)$ is easy to compute ...
- ... since we use Dirichlet distributions for $P(\Theta|T, X)$ (and other standard assumptions).

Comparing class probabilities in an easy case

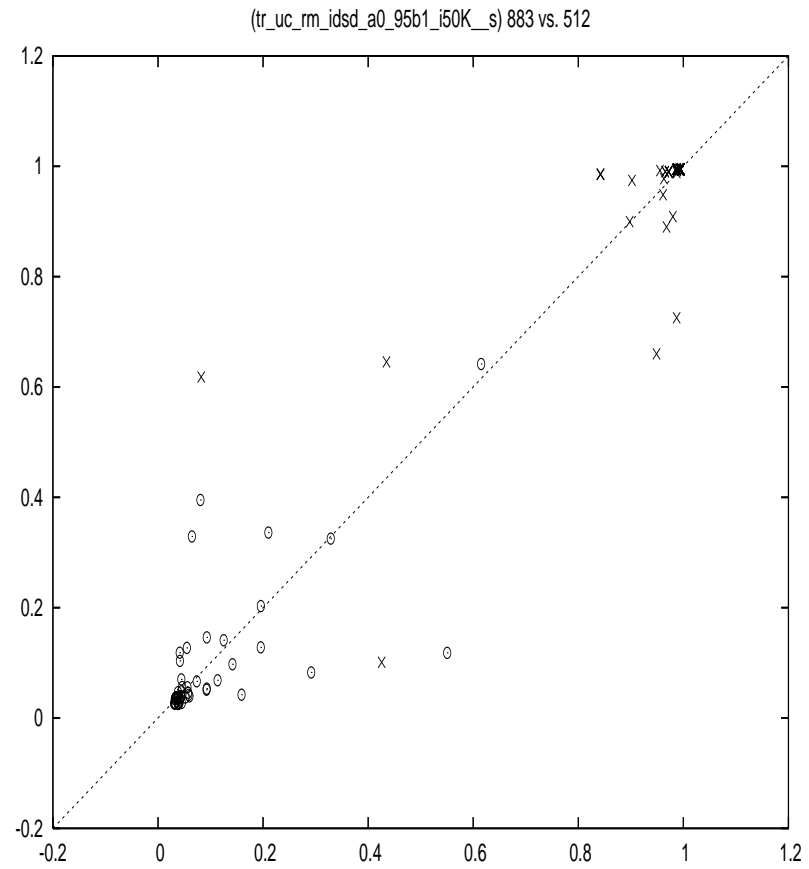
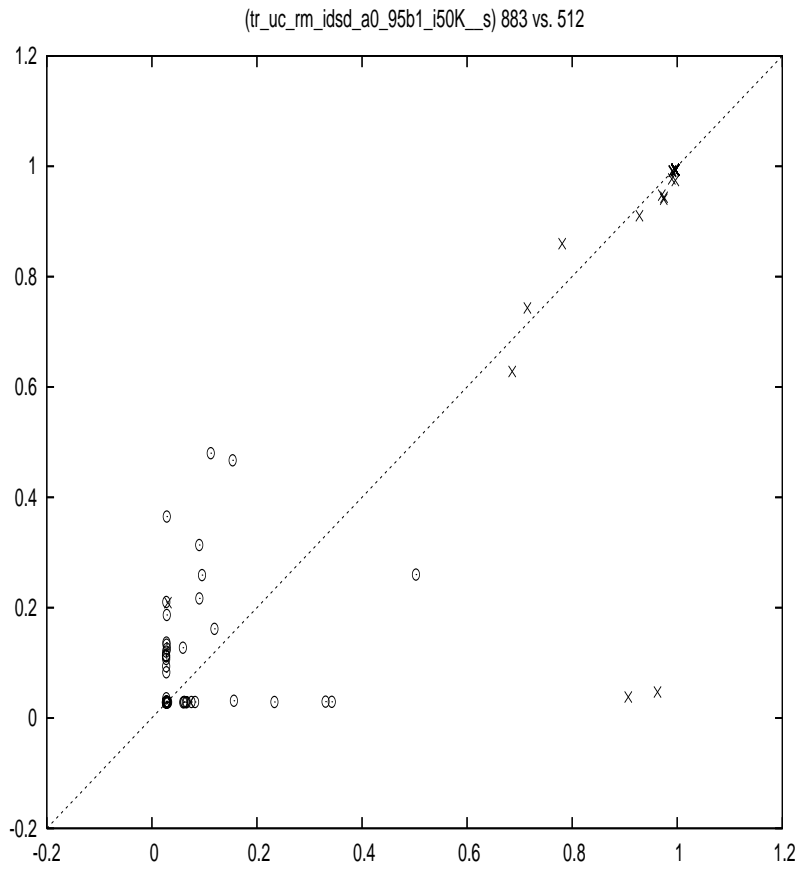
1_i50K__s) 883 vs. 209



Dataset=K, iterations=50,000, tempering=FALSE

Bayesian C&RT

BCW: 50K, Temp=F vs 50K, Temp=T



Bayesian C&RT