Preliminary Analysis of MKL

Flat Maxima, Diversity and Fisher Information

Theo Damoulas
Institute for Computational Sustainability
Computing and Information Science
Cornell University
Ithaca, NY, USA

Mark A. Girolami
Inference Research Group
Dept. of Computing Science & Dept. of Statistics
University of Glasgow
Scotland, UK

Simon Rogers
Inference Research Group
Department of Computing Science
University of Glasgow
Scotland, UK

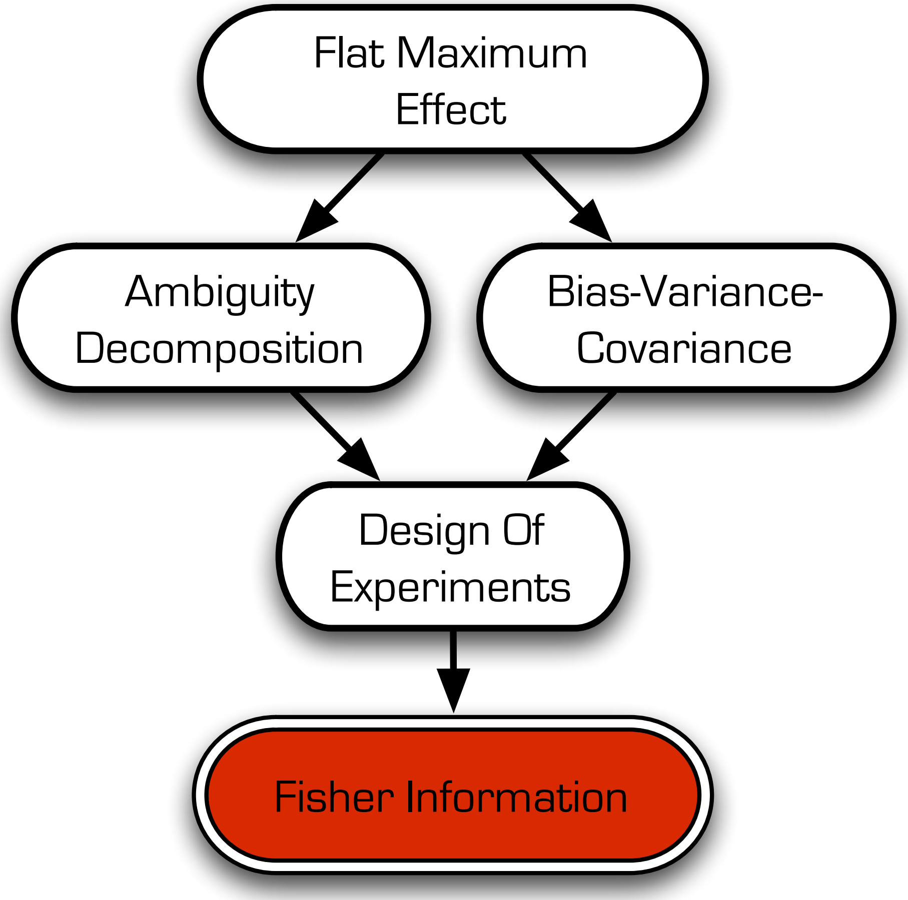
Understanding Multiple Kernel Learning Methods
NIPS Workshop, Whistler, December 2009

CORNELL UNIVERSITY FOUNDED A.D. 1865

University of Glasgow

Preliminary Analysis of MKL

Flat Maxima, Diversity and Fisher Information



Flat Maximum Effect

- D. J. Hand (Statistical Science, 21(1) p. 1-15, 2006) - Linear regression

$$w = \sum_{i=1}^{d} w_i x_i \qquad u = \sum_{i=1}^{d} u_i x_i \qquad w_i, u_i \geq 0 \quad \sum_{i=1}^{d} w_i = \sum_{i=1}^{d} u_i = 1$$

$$\rho(u, w) \geq \sum_{i,j=1}^{d} u_i w_j \rho(x_i, x_j)$$

"Often quite large deviations from the optimal set of weights will yield predictive performance not substantially worse than the optimal weights"

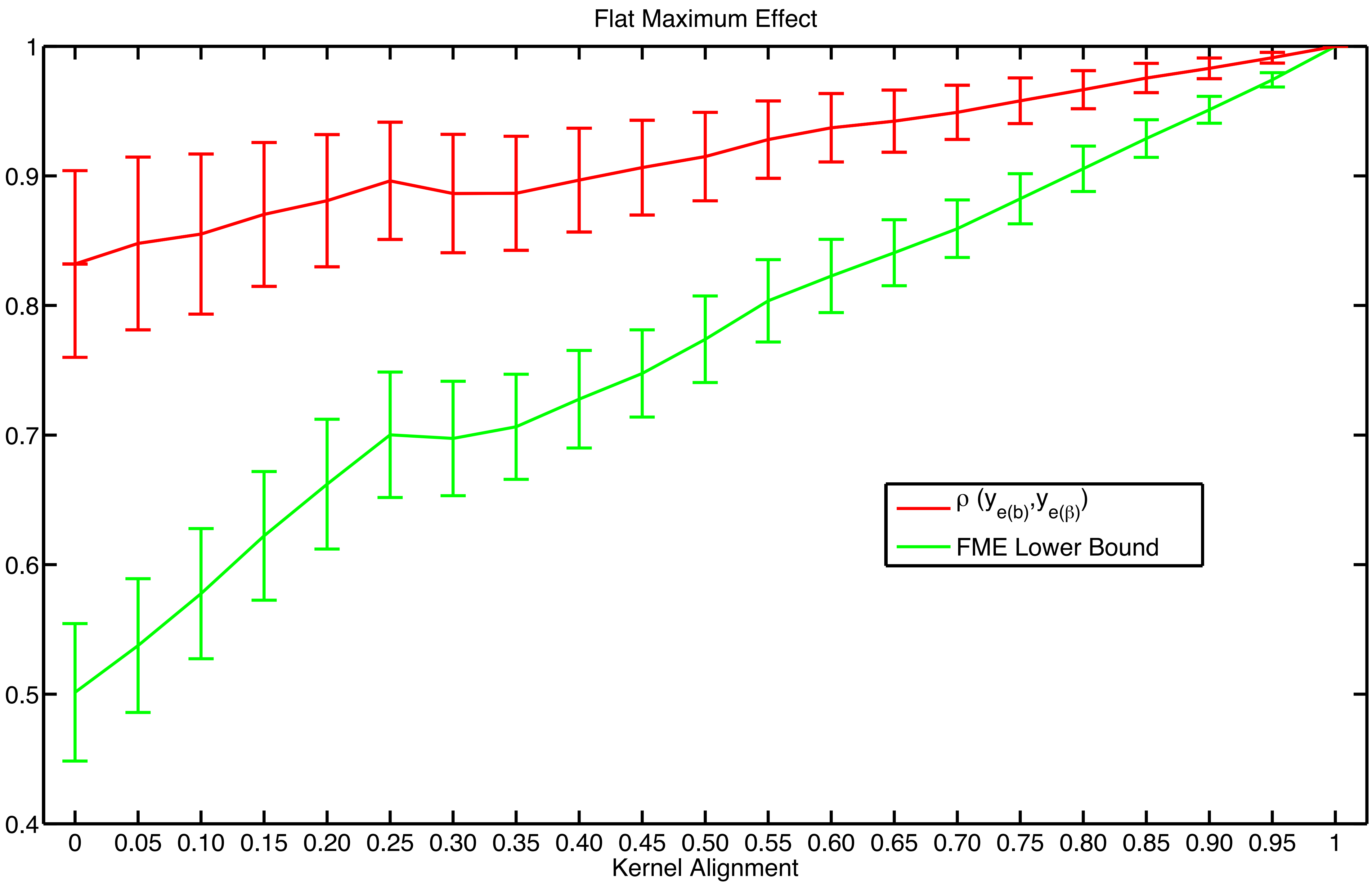
$$\mathbf{y}_{e_{(\boldsymbol{b})}} = \sum_{j=1}^{N} w_j \sum_{s=1}^{S} b_s k_s(\mathbf{x}_i, \mathbf{x}_j) = \sum_{s=1}^{S} b_s \mathbf{y}_s \qquad \mathbf{y}_{e_{(\boldsymbol{\beta})}} = \sum_{j=1}^{N} w_j \sum_{\sigma=1}^{S} \beta_\sigma k_\sigma(\mathbf{x}_i, \mathbf{x}_j) = \sum_{\sigma=1}^{S} \beta_\sigma \mathbf{y}_\sigma$$

Flat Maximum Lower Bound

$$\boxed{\rho(\mathbf{y}_{e_{(\boldsymbol{b})}}, \mathbf{y}_{e_{(\boldsymbol{\beta})}}) \geq \sum_{s,\sigma=1}^{S} b_s \beta_\sigma \rho(\mathbf{y}_s, \mathbf{y}_\sigma)}$$

The correlation between any two weighted kernel combination responses is lower bounded by a function of the correlation between base kernel responses.

- Sample Kernels - Sample Kernel Combination Parameters - Sample Regression Coefficients



- Ensemble response correlation increases as kernel alignment increases.
- Little or no benefit in parameterized (global) kernel combinations for highly aligned base kernels.

Decompositions of the Loss

Ambiguity

$$\underbrace{(\mathbf{y}_e-\mathbf{y})^T(\mathbf{y}_e-\mathbf{y})}_{\text{Composite Error}} = \underbrace{\sum_{s=1}^{S}\beta_s(\mathbf{y}_s-\mathbf{y})^T(\mathbf{y}_s-\mathbf{y})}_{\text{Weighted Ind. Error}} - \underbrace{\sum_{s=1}^{S}\beta_s(\mathbf{y}_s-\mathbf{y}_e)^T(\mathbf{y}_s-\mathbf{y}_e)}_{\text{Ambiguity}}$$

Brown & Wyatt
ICML '03

- Need for accurate (Weighted Ind. Error) but diverse (Ambiguity) base kernels.

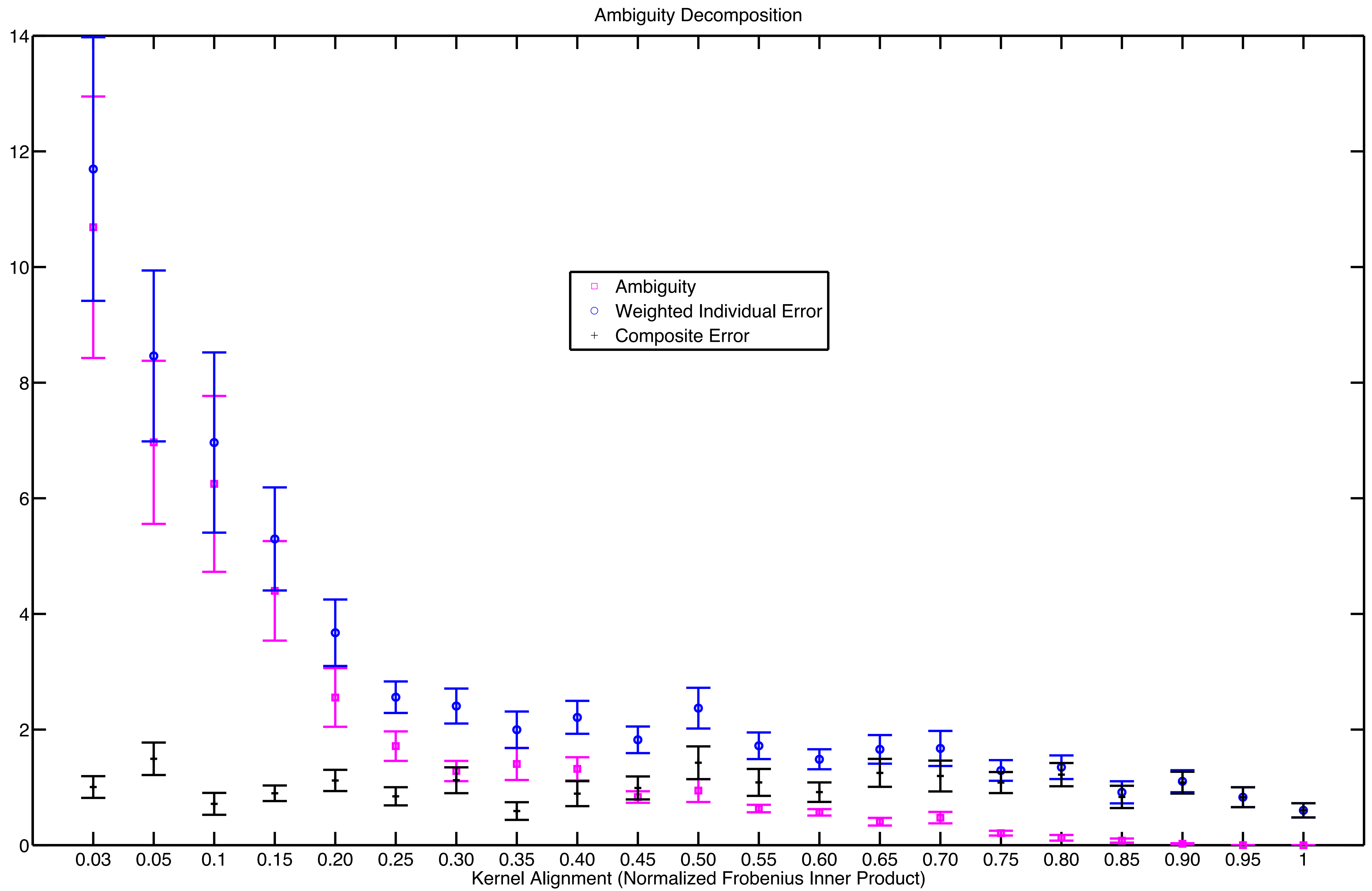
Bias
vs
Variance
vs
Covariance

$$\mathbb{E}\left\{(\mathbf{y}_e-\mathbf{y})^{\mathrm{T}}((\mathbf{y}_e-\mathbf{y})\right\} = \sum_{s,\sigma=1}^{S}\beta_s\beta_\sigma\underbrace{(\mathbb{E}\{\mathbf{y}_s\}-\mathbf{y})^{\mathrm{T}}(\mathbb{E}\{\mathbf{y}_\sigma\}-\mathbf{y})}_{\mathrm{Bias}^{\mathrm{T}}\mathrm{Bias}}$$
$$+\sum_{s=1}^{S}\beta_s^2\underbrace{\mathbb{E}\left\{(\mathbf{y}_s-\mathbb{E}\{\mathbf{y}_s\})^{\mathrm{T}}(\mathbf{y}_s-\mathbb{E}\{\mathbf{y}_s\})\right\}}_{\text{Variance}} + \sum_{s=1,\sigma\neq s}^{S}\beta_s\beta_\sigma\underbrace{\mathbb{E}\left\{(\mathbf{y}_s-\mathbb{E}\{\mathbf{y}_s\})^{\mathrm{T}}(\mathbf{y}_\sigma-\mathbb{E}\{\mathbf{y}_\sigma\})\right\}}_{\text{Covariance}}$$

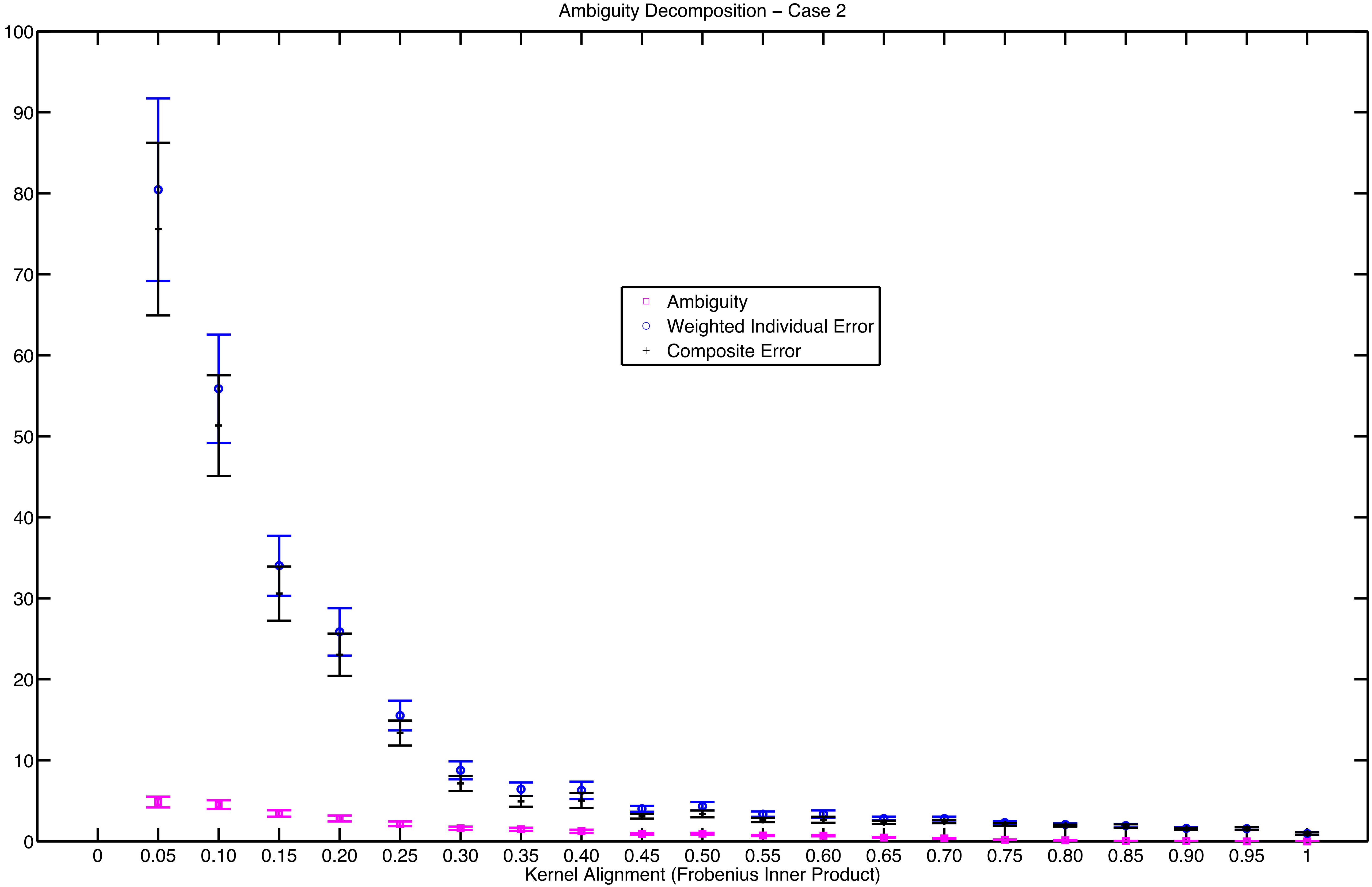
Ueda & Nakano
ICNN '96

- Fitting (Bias) versus Generalisation (Variance) versus Diversity (Covariance).

- Sample data from ensemble response.
- Quantifying the need for accurate and diverse base kernels.



- Sampling data from single base kernel response.
- Varying composite error - high individual error - diverse but not accurate.



Design of Experiments

- MKL as an Experimental Design problem [R. A. Fisher, The Design of Experiments,1935]
- Maximize the information offered for the model parameters with respect to the evidence observed

$$\mathbf{F}(\boldsymbol{\theta}) = -\mathbb{E}\left\{\frac{\partial^2 \mathcal{L}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right\}$$

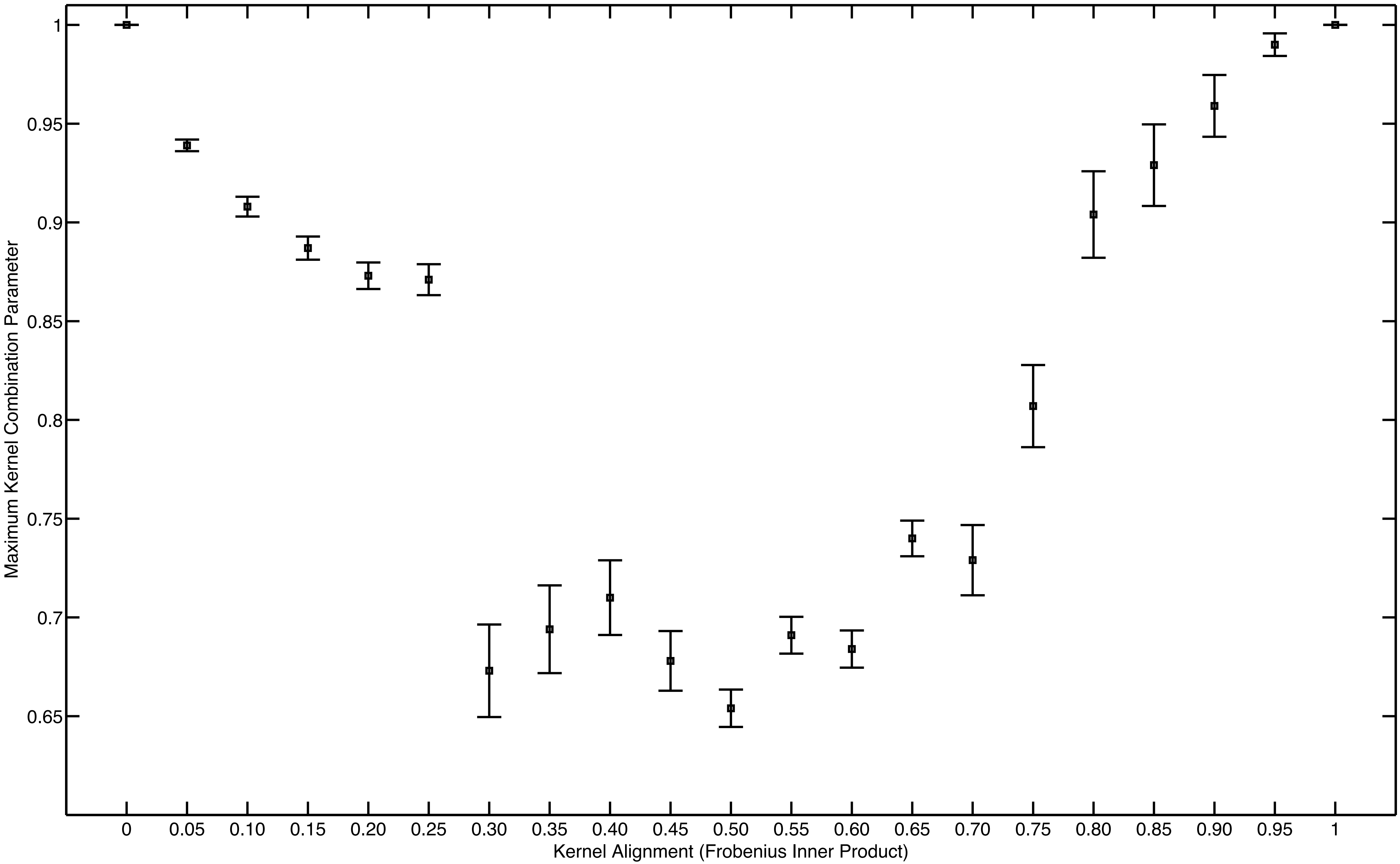
$$\mathcal{L} = \left(\mathbf{y}^{\mathrm{T}}\sum_{s=1}^{S}\beta_s\mathbf{K}_s\mathbf{w} - \frac{1}{2}\mathbf{w}^{\mathrm{T}}\sum_{s,\sigma=1}^{S}\beta_s\beta_\sigma\mathbf{K}_s\mathbf{K}_\sigma\mathbf{w}\right)$$

$$\mathbf{F}_{s\sigma}(\boldsymbol{\beta}) = \mathbf{w}^{\mathrm{T}}\mathbf{K}_s^{\mathrm{T}}\mathbf{K}_\sigma\mathbf{w} \qquad \mathbf{F}(\mathbf{w}) = \mathbf{K}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{K}_{\boldsymbol{\beta}}$$

- Information Optimality Criteria: Maximize Information by minimizing variance of estimator.
- D-optimality :

$$\beta \leftarrow \underset{\beta}{\operatorname{argmin}}\left\{-2\log|\mathbf{K}_\beta|\right\}$$

- Maximum value of kernel combination parameter while varying alignment.
- Fisher MKL approach and an area of “learning”.



Conclusions

- FME and a lower bound for correlation of MKL responses - flat maxima.
- Ambiguity and BVC decompositions - accuracy & diversity.
- DoE approach for simple linear regression MKL case - Fisher information.

What's next

- Fisher MKL for Classification.
- Analyzing the effect of sparsity (prior-regularization) and "localized" MKL?

Code

- www.dcs.gla.ac.uk/inference/pMKL (Variational Bayes & Sparse models)

Preliminary Analysis of MKL

Flat Maxima, Diversity and Fisher Information

Thank you for your attention

Theo Damoulas
Institute for Computational Sustainability
Computing and Information Science
Cornell University
Ithaca, NY, USA

Mark A. Girolami
Inference Research Group
Dept. of Computing Science & Dept. of Statistics
University of Glasgow
Scotland, UK

Simon Rogers
Inference Research Group
Department of Computing Science
University of Glasgow
Scotland, UK

Understanding Multiple Kernel Learning Methods
NIPS Workshop, Whistler, December 2009

University of Glasgow