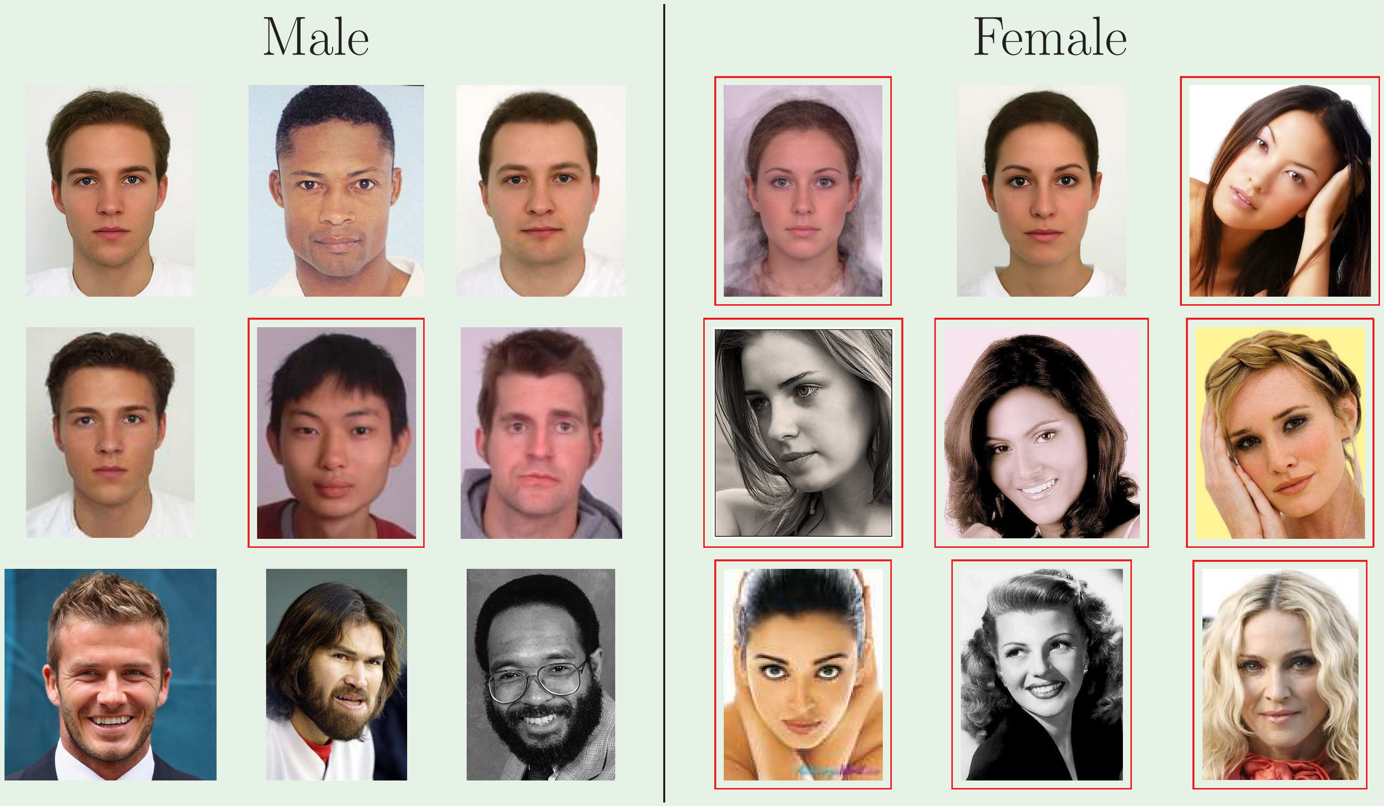
A Permutation Approach to Validation

M. Magdon-Ismail, Konstantin Mertsalov

Rensselaer

Siam Conference on Data Mining (SDM)
April 30 2010

Example: Learning Male Vs. Female Faces



Learned rule: "**roundish face or long hair is female**"

$e_{\text{in}} = \frac{2}{18} \approx 11\%$

$e_{\text{out}} = ??$

It has been known since the early days that $e_{\text{in}} \ll e_{\text{out}}$.

[Larson, 1931; Wherry, 1931, 1951; Katzell, 1951; Cureton, 1951; Mosier, 1951; Stone, 1974]

Generalization Error

$$e_{\text{gen}} = e_{\text{out}} - e_{\text{in}}$$

- **Statistical Methods**: FPE; GCV; Covariance penalties; etc.

  [Akaike, 1974; Craven and Wahba, 1979; Efron, 2004; Wang and Shen, 2006].

  – Generally assume a well specified model.

- **Uniform Bounds**:

  – Distribution independent: VC [Vapnik and Chervonenkis, 1971].

  – Data dependent: Maximum discrepancy; Rademacher-style; margin bounds.

  [Bartlett *et al.*, 2002; Bartlett and Mendelson, 2002; Fromont, 2007; Kääriäinen and Elomaa, 2003; Koltchinskii, 2001; Koltchinskii and Panchenko, 2000; Lozano, 2000; Lugosi and Nobel, 1999; Massart, 2000; Shawe-Taylor *et al.*, 1998].

- **Sampling methods**: Leave-K-out cross validation. [Stone, 1974]

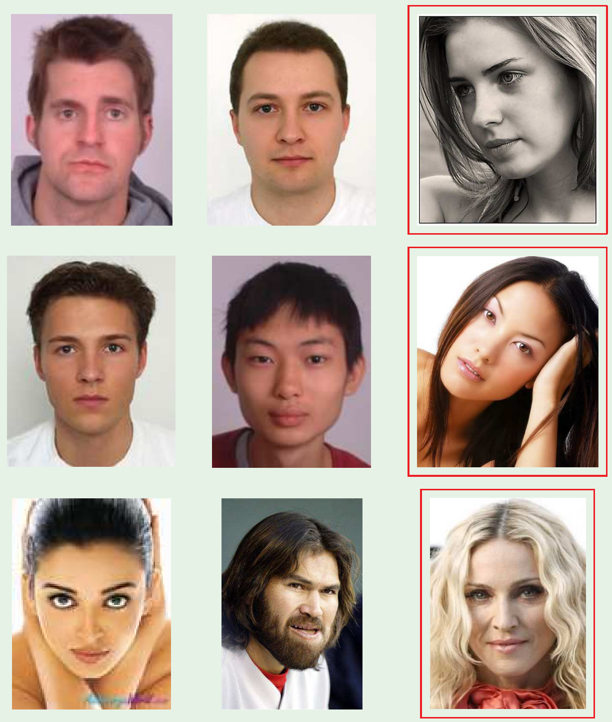
- **Permutation Methods**: have been used as tests of significance for model selection.

  [Golland *et al.*, 2005; Wiklund *et al.*, 2007]

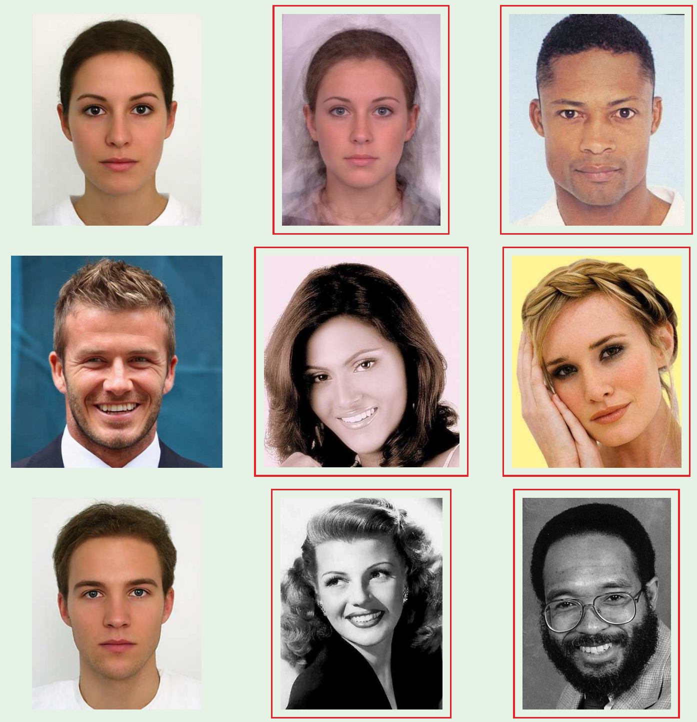
We will present a permutation method for validation – estimation of e_{gen}.

An "Artificial" Permuted Problem π

"Male" permuted data



"Female" permuted data



Learned rule: "**dark skin or long hair is female**"

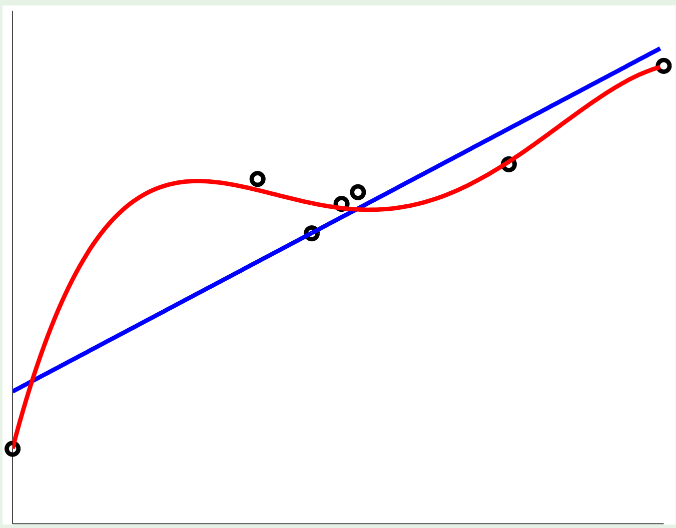
$e_{\text{in}}^{\boldsymbol{\pi}} = \frac{6}{18} \approx 33\%$

$e_{\text{out}}^{\boldsymbol{\pi}} = 50\%$

$\widehat{e}_{\text{gen}} \approx 17\% \longleftarrow$ Use this to estimate $\widehat{e}_{\text{out}} = e_{\text{in}} + \widehat{e}_{\text{gen}} \approx 28\%$.

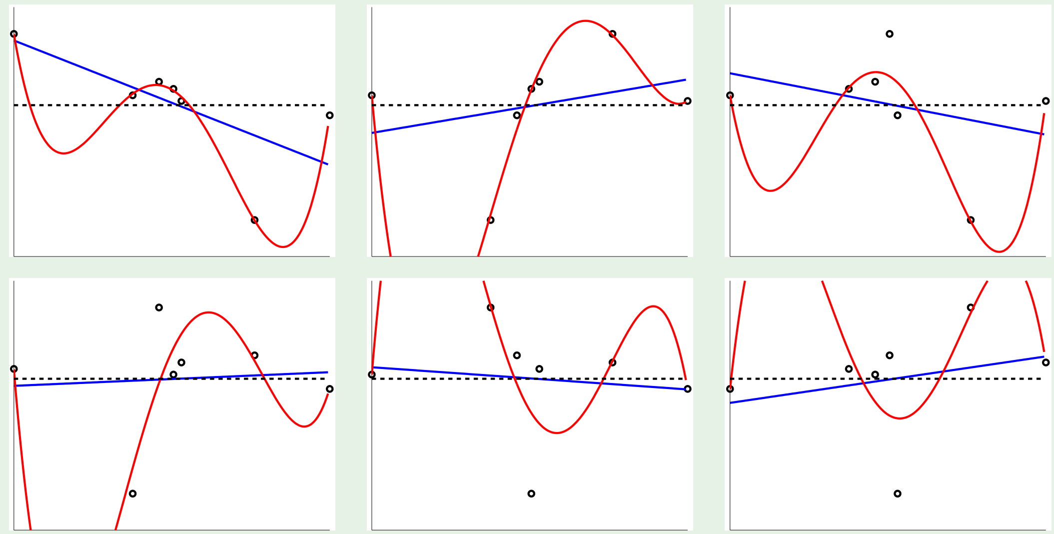
Permutation Method for Regression

Real Data



Linear Fit	Quartic Fit
$e_{\text{in}} = 0.02$	$e_{\text{in}} = 0.002$
$e_{\text{out}} = 0.11$	$e_{\text{out}} = 0.256$
$e_{\text{gen}} = 0.08$	$e_{\text{gen}} = 0.254$
$\widehat{e}_{\text{out}} = 0.07$	$\widehat{e}_{\text{out}} = 0.192$

Permuted Data



Linear Fit	Quartic Fit
$\text{average}(e_{\text{in}}{}^{\pi}) = 0.12$	$\text{average}(e_{\text{in}}{}^{\pi}) = 0.05$
$\text{average}(e_{\text{out}}{}^{\pi}) = 0.17$	$\text{average}(e_{\text{out}}{}^{\pi}) = 0.24$
$\text{average}(\widehat{e}_{\text{gen}}) = 0.05$	$\text{average}(\widehat{e}_{\text{gen}}) = 0.19$

The Permutation Method For Validation

1. Fit the real data to obtain $e_{\text{in}}(g)$.
2. Permute the y values using permutation $\boldsymbol{\pi}$.
   (a) Fit the permuted data to obtain $g^{\boldsymbol{\pi}}$
   (b) Compute the generalization error on the artificial permuted problem.

   Theorem 1. $e^{\boldsymbol{\pi}}_{\text{out}}(g^{\boldsymbol{\pi}}) = s_y^2 + \frac{1}{n}\sum_{i=1}^{n}(g^{\boldsymbol{\pi}}(x_i) - \bar{y})^2.$

   Theorem 2. $e^{\boldsymbol{\pi}}_{\text{gen}}(g^{\boldsymbol{\pi}}) = \frac{2}{n}\sum_{i=1}^{n}(y_{\pi_i} - \bar{y})g^{\boldsymbol{\pi}}(x_i)$

   (Twice the (spurious) correlation between $g^{\boldsymbol{\pi}}$ and $y^{\boldsymbol{\pi}}$.)
3. Repeat (say 100 times) to get an average($\widehat{e}_{\text{gen}}$).
4. Estimate the out-sample error

$$\widehat{e}_{\text{out}} = e_{\text{in}} + \widehat{e}_{\text{gen}}.$$

Example Linear Ridge Regression

$$g(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

Construct $\mathbf{w}_{\text{in}}$ to minimize $e_{\text{in}}(\mathbf{w}) + \lambda \mathbf{w}^{\mathrm{T}}\mathbf{w}$. The in-sample predictions are

$$\hat{\mathbf{y}} = \mathrm{S}(\lambda)\mathbf{y},$$

where, $\mathrm{S}(\lambda) = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda \mathrm{I})^{-1}\mathrm{X}^{\mathrm{T}}$.

Theorem 3.

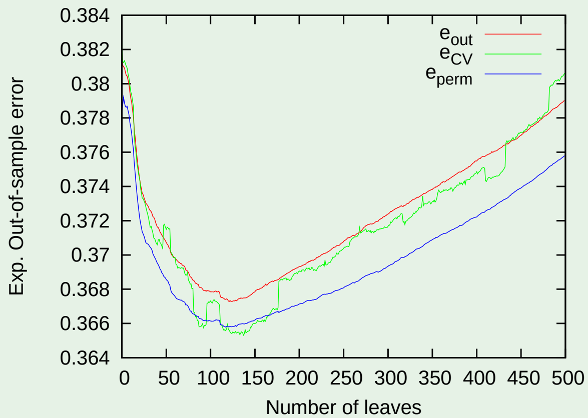
$$\widehat{e}_{\text{out}}(g) = e_{\text{in}}(g) + \frac{2\hat{\sigma}_y^2}{n}\left(\text{trace}(\mathrm{S}) - \frac{\mathbf{1}^{\mathrm{T}}\mathrm{S}\mathbf{1}}{n}\right).$$

When $(\lambda = 0)$, S is a projection matrix:

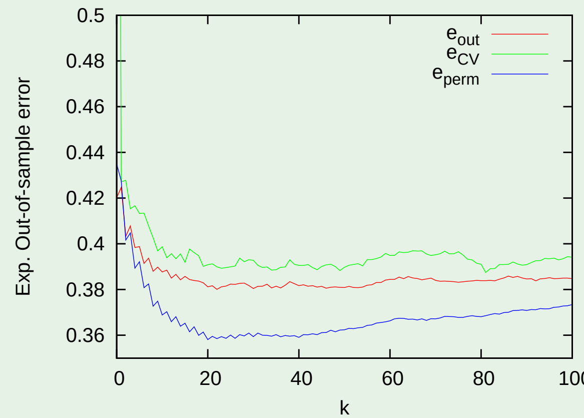
$$\widehat{e}_{\text{out}} = e_{\text{in}} + \frac{2\hat{\sigma}_y^2 d}{n}.$$

(An Akaike FPE-type estimator; $\hat{\sigma}_y^2 = \frac{n}{n-1}s_y^2$, the unbiased estimate of the y-variance.)

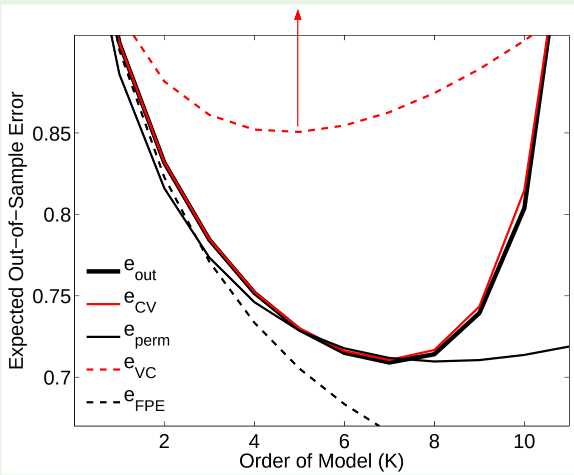
Validation Results



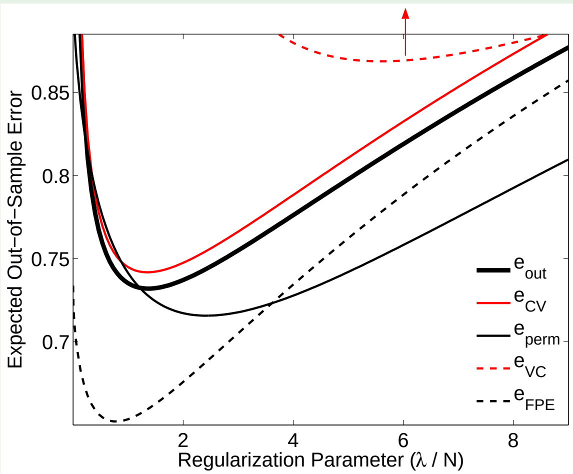
(a) LOO-CV vs. Permutation (DT)



(b) LOO-CV vs. Permutation (k-NN).



(a) Different Polynomial Order.



(b) Different Regularization Parameter.

Model Selection – Simulated Setting

Validation Estimate	Order Selection		λ Selection		
			Unregularized		Regularized
	Regret	Avg Order	Regret	Avg. $\frac{\lambda}{N}$	Regret
LOO-CV	540	9.29	18.8	23.1	0.44
Perm.	**185**	**7.21**	5.96	9.57	**0.39**
VC	508	5.56	**3.50**	**125**	0.42
FPE	9560	11.42	51.3	18.1	0.87

Noise(%)	LOO-CV	Perm.	Rad.
5	0.30	**0.28**	**0.28**
10	0.28	**0.27**	**0.27**
15	0.28	**0.25**	**0.25**
20	0.28	**0.26**	**0.26**
25	0.26	**0.25**	**0.25**
30	**0.24**	**0.24**	**0.24**

Model Selection – Real Data

Data	Decision Trees			k-Nearest Neighbor		
	LOO-CV	Perm.	Rad.	LOO-CV	Perm.	Rad.
Abalone	0.05	**0.02**	**0.02**	0.04	0.04	0.04
Ionosphere	0.17	**0.16**	0.17	**0.17**	0.70	0.83
M.Mass	0.09	**0.05**	**0.05**	**0.09**	0.11	0.11
Parkinsons	**0.24**	0.34	0.41	**0.25**	0.33	0.43
Pima Diabetes	0.09	**0.07**	**0.07**	**0.11**	**0.11**	0.14
Spambase	0.07	**0.06**	0.07	**0.19**	0.43	0.55
Transfusion	0.10	**0.08**	0.09	**0.09**	0.12	0.19
WDBC	**0.20**	0.23	0.34	**0.21**	0.34	0.51
Diffusion	0.04	0.03	**0.02**	0.04	0.06	**0.03**
Simulated	0.16	**0.15**	**0.15**	0.21	0.21	0.21

Learning episodes limited to 10

Data	Decision Trees				k-Nearest Neighbor		
	LOO-CV	10-fold	Perm.	Rad.	LOO-CV	Perm.	Rad.
Abalone	0.12	0.13	**0.02**	**0.02**	0.24	**0.09**	0.12
Ionosphere	0.24	0.21	**0.18**	0.19	**0.49**	0.75	0.84
M.Mass	0.23	0.13	**0.06**	**0.06**	0.15	**0.11**	0.12
Parkinsons	**0.25**	0.31	0.34	0.40	0.34	**0.32**	0.44
Pima Diabetes	0.18	0.18	**0.07**	**0.07**	0.16	0.12	0.15
Spambase	0.28	0.09	**0.07**	**0.07**	0.44	**0.43**	0.54
Transfusion	0.19	0.13	**0.08**	0.09	0.17	**0.12**	0.19
WDBC	0.31	0.40	**0.24**	0.37	0.55	**0.33**	0.50
Diffusion	0.13	0.04	0.03	**0.02**	0.09	0.06	**0.04**

What Have We Learned?

- To estimate e_{out}: hard to beat LOO-CV (in expectation).
- Model selection: need good estimate, *but also stable*.
- VC – ultra stable, very conservative.
- LOO-CV – very unstable, in general good, but can be a disaster.
- Permutation Method – Good blend.
  - To have low $\widehat{e}_{\text{gen}}$, the method must generalize well on random permutations which have similar structure to the data. This induces stability.
  - Seems to be better than Rademacher, which is of a similar flavor: the permutation preserves more of the structure of the data, while at the same time being stable.

. . . And Now the Theory: Permutation Complexity

Permutation Complexity

$$\mathcal{P}_{\text{in}}(\mathcal{H}|D) = \mathbb{E}_{\boldsymbol{\pi}}\left[\max_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} y_{\pi_i} h(x_i)\right].$$

We consider random permutations $\boldsymbol{\pi}$ of the y values.

Some function in your hypothesis set achieves a maximum (spurious) correlation with this random permutation.

The expected value of this spurious correlation is the *permutation complexity*.

- *data dependent.*
- can be computed by empirical error minimization.

[Rademacher complexity is similar except that it chooses y_i independently and uniformly in $\{\pm 1\}$.]

Permutation Complexity Uniform Bound

Theorem 4.

$$\begin{aligned} e_{\text{out}}(g) &\le e_{\text{in}}(g) + 4\mathcal{P}_{\text{in}}(\mathcal{H}|D) + O\left(\sqrt{\tfrac{1}{n}\ln\tfrac{1}{\delta}}\right), \\ &\overset{(*)}{=} e_{\text{in}}(g) + 2\widehat{e}_{\text{gen}}(\mathcal{H}|D) + 4\bar{y}\,\mathbb{E}_{\boldsymbol{\pi}}\left[\bar{g}^{\boldsymbol{\pi}}\right] + O\left(\sqrt{\tfrac{1}{n}\ln\tfrac{1}{\delta}}\right). \end{aligned}$$

$(*)$ is for empirical risk minimization (ERM).

Up to a small "bias term", $\widehat{e}_{\text{gen}}$ bounds e_{out} (for ERM).

The bound is uniform, and data dependent.

Practical "consequence": we are "justified" in using the permutation estimate.

Proof

- We now have tools for *i.i.d.* sampling: McDiarmid's Inequality [McDiarmid, 1989].
- The main difficulty: permutation sampling is *not* independent.
- The insight is to use multiple *ghost samples* to "unfold" this dependence.
- …one still has to go through a few technical details, but then you have it.

Wrapping Up

- The permutation estimate is easy to compute numerically - all you do is run the algorithm on randomly permuted data.
- Can be used for classification or regression.
- In some cases (linear ridge regression), can get analytical form.
- Achieves a good blend (practically) between the conservative VC bound and the highly unstable LOO-CV.
- Similar but slightly superior (in practice) to Rademacher penalties.
- ...its only the begining.

Thank You! Questions?

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Aut. Cont.*, **19**, 716–723.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, **3**, 463–482.

Bartlett, P. L., Boucheron, S., and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning*, **48**, 85–113.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377–403.

Cureton, E. E. (1951). Symposium: The need and means of cross-validation: II approximate linear restraints and best predictor weights. *Education and Psychology Measurement*, **11**, 12–15.

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, **99**(467), 619–632.

Fromont, M. (2007). Model selection by bootstrap penalization for classification. *Machine Learning*, **66**(2-3), 165–207.

Golland, P., Liang, F., Mukherjee, S., and Panchenko, D. (2005). Permutation tests for classification. *Learning Theory*, pages 501–515.

Kääriäinen, M. and Elomaa, T. (2003). Rademacher penalization over decision tree prunings. In *In Proc. 14th European Conference on Machine Learning*, pages 193–204.

Katzell, R. A. (1951). Symposium: The need and means of cross-validation: III cross validation of item analyses. *Education and Psychology Measurement*, **11**, 16–22.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, **47**(5), 1902–1914.

Koltchinskii, V. and Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. In E. Gine, D. Mason, and J. Wellner, editors, *High Dimensional Prob. II*, volume 47, pages 443–459.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Education Psychology*, **22**, 45–55.

Lozano, F. (2000). Model selection using rademacher penalization. In *Proc. 2nd ICSC Symp. on Neural Comp.*

Lugosi, G. and Nobel, A. (1999). Adaptive model selection using empirical complexities. *Annals of Statistics*, **27**, 1830–1864.

Massart, P. (2000). Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciencies de Toulouse*, **X**, 245–303.

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press.

Mosier, C. I. (1951). Symposium: The need and means of cross-validation: I problem and designs of cross validation. *Education and Psychology Measurement*, **11**, 5–11.

Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., and Anthony, M. (1998). Structural risk minimization over data dependent hierarchies. *IEEE Transactions on Information Theory*, **44**, 1926–1940.

Stone, M. (1974). Cross validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, **36**(2), 111–147.

Vapnik, V. N. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their pr obabilities. *Theory of Probability and its Applications*, **16**, 264–280.

Wang, J. and Shen, X. (2006). Estimation of generalization error: random and fixed inputs. *Statistica Sinica*, **16**, 569–588.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the multiple correlation coefficient. *Annals of Mathematical Statistics*, **2**, 440–457.

Wherry, R. J. (1951). Symposium: The need and means of cross-validation: III comparison of cross validation with statistical inference of betas and multiple r from a single sample. *Education and Psychology Measurement*, **11**, 23–28.

Wiklund, S., Nilsson, D., Eriksson, L., Sjostrom, M., Wold, S., and Faber, K. (2007). A randomization test for pls component selection. *Journal of Chemometrics*, **21(10,11)**.