ON OPTIMAL DATA SPLIT FOR GENERALIZATION ESTIMATION AND MODEL SELECTION

Jan Larsen and Cyril Goutte
Department of Mathematical Modeling, Building 321
Technical University of Denmark, DK-2800 Lyngby, Denmark
E-mail: jl,cg@imm.dtu.dk, Web: eivind.imm.dtu.dk

OVERVIEW

Modeling with flexible models, such as neural networks, requires careful control of the model complexity and generalization ability of the resulting model. Whereas general asymptotic estimators of generalization ability have been developed over recent years (e.g., [9]), it is widely acknowledged that in most modeling scenarios there isn't sufficient data available to reliably use these estimators for assessing generalization, or select/optimize models. As a consequence, one resorts to resampling techniques like cross-validation [3, 8, 14], jackknife or bootstrap [2]. In this paper, we address a crucial problem of cross-validation estimators: how to split the data into various sets.

The set $\mathcal{D}$ of all available data is usually split into two parts: the design set $\mathcal{E}$ and the test set $\mathcal{F}$. The test set is exclusively reserved to a final assessment of the model which has been designed on $\mathcal{E}$ (using e.g., optimization and model selection). This usually requires that the design set in turn is split in two parts: training set $\mathcal{T}$ and validation set $\mathcal{V}$. The objective of the *design/test* split is to both obtain a model with high generalization ability and to assess the generalization error reliably. The second split is the *training/validation* split of the design set. Model parameters are trained on the training data, while the validation set provides an estimator of generalization error used to e.g., choose between alternative models or optimize additional (hyper) parameters such as regularization or robustness parameters [10, 12]. The aim is to select the split so that the generalization ability of the resulting model is as high as possible.

This paper is concerned with studying the very different behavior of the two data splits using *hold-out cross-validation*, *K-fold cross-validation* [3, 14] and *randomized permutation cross-validation*[1] [1], [13, p. 309]. First we describe the theoretical basics of various cross-validation techniques with the purpose of reliably estimating the generalization error and optimizing the

[1]Also called monte-carlo cross-validation or repeated learning-testing methods.

model structure. The next section deals with the simple problem of estimating a single location parameter. This problem is tractable as non-asymptotic theoretical analysis is possible, whereas mainly asymptotic analysis and simulation studies are viable for the more complex AR-models and neural networks discussed in the subsequent sections.

TRAINING AND GENERALIZATION

Suppose that our model $\mathcal{M}$ (e.g., neural network) is described by the function $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{w})$ where $\boldsymbol{x}$ is the input vector and $\boldsymbol{w}$ is the vector of parameters (or weights) with dimensionality m. The objective is to use the model for approximating the true conditional input-output distribution $p(\boldsymbol{y}|\boldsymbol{x})$, or some moments thereof. For regression and signal processing problems we normally model the conditional expectation $E\{\boldsymbol{y}|\boldsymbol{x}\}$. Define the training set $\mathcal{T} = \{\boldsymbol{x}(k);\boldsymbol{y}(k)\}_{k=1}^{N_\mathcal{T}}$ of $N_\mathcal{T}$ input-output examples sampled from the unknown but fixed joint input-output probability density $p(\boldsymbol{x},\boldsymbol{y})$. The model is trained by minimizing a cost function $C_\mathcal{T}(\boldsymbol{w})$, usually the sum of a loss function (or training error), $S_\mathcal{T}(\boldsymbol{w})$, and a regularization term $R(\boldsymbol{w},\boldsymbol{\kappa})$ parameterized by a set of regularization parameters $\boldsymbol{\kappa}$:

$$C_\mathcal{T}(\boldsymbol{w}) = S_\mathcal{T}(\boldsymbol{w}) + R(\boldsymbol{w},\boldsymbol{\kappa}) = \frac{1}{N_\mathcal{T}}\sum_{k\in\mathcal{T}} \ell\left(\boldsymbol{y}(k),\widehat{\boldsymbol{y}}(k)\right) + R(\boldsymbol{w},\boldsymbol{\kappa}) \qquad (1)$$

where $\ell(\cdot)$ measures the cost of estimating the output $\boldsymbol{y}(k)$ with the model prediction $\widehat{\boldsymbol{y}}(k) = \boldsymbol{f}(\boldsymbol{x}(k);\boldsymbol{w})$, e.g., log-likelihood loss or the simple squared error loss function $\ell(\boldsymbol{y},\widehat{\boldsymbol{y}}) = \|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|^2$. Training provides the estimated weight vector $\widehat{\boldsymbol{w}} = \operatorname{argmin}_{\boldsymbol{w}} C_\mathcal{T}(\boldsymbol{w})$. *Generalization error* is defined as the expected loss on a future independent sample $(\boldsymbol{x},\boldsymbol{y})$,

$$G(\widehat{\boldsymbol{w}}) = E_{\boldsymbol{x},\boldsymbol{y}}\{\ell(\boldsymbol{y},\widehat{\boldsymbol{y}})\} = \int \ell(\boldsymbol{y},\widehat{\boldsymbol{y}})\, p(\boldsymbol{x},\boldsymbol{y})\, d\boldsymbol{x}d\boldsymbol{y} \qquad (2)$$

The *average generalization error* Γ is defined by averaging $G(\widehat{\boldsymbol{w}})$ over all possible training sets (see also [11]):

$$\Gamma = E_\mathcal{T}\{G(\widehat{\boldsymbol{w}})\} = \int G(\widehat{\boldsymbol{w}})\, p(\mathcal{T})\, d\mathcal{T}. \qquad (3)$$

Optimization of the model structure, including e.g., regularization parameters [12], is done by minimizing an empirical estimate of the generalization error based on the validation data. Finally, the test data provides an unbiased empirical estimate of the generalization error of the resulting model.

Generalization Assessment

Given a data set $\mathcal{D} = \{\boldsymbol{x}(k);\boldsymbol{y}(k)\}_{k=1}^{N}$ of N *independent* input-output examples, let us first consider the split of the data into the design and test sets, denoted by $\mathcal{E}$ and $\mathcal{F}$ respectively. The purpose is to design a model achieving maximum generalization performance and assess this performance as reliably as possible. We consider three methods:

Hold-Out Cross-Validation (HO). An empirical estimate of (2) is obtained by splitting the data once into design and test sets. Define γ as the split ratio leaving $N_{\mathcal{F}} = \gamma N$ for testing and $N_{\mathcal{E}} = (1-\gamma)N$ for design[2]. The HO estimate of generalization error for model $\widehat{\boldsymbol{y}}$ (with weights $\widehat{\boldsymbol{w}}$) designed on $\mathcal{E}$ is given by

$$\widehat{G}_{\mathrm{HO}}(\widehat{\boldsymbol{w}}) = \frac{1}{N_{\mathcal{F}}} \sum_{k \in \mathcal{F}} \ell(\boldsymbol{y}(k), \widehat{\boldsymbol{y}}(k)). \tag{4}$$

The quality of the estimate is evaluated by considering the mean squared error:

$$\begin{aligned} \mathrm{MSE}_{\mathrm{HO}}(\gamma) &= E_{\mathcal{D}}\left\{ \left(\widehat{G}_{\mathrm{HO}}(\widehat{\boldsymbol{w}}) - G(\boldsymbol{w}^*) \right)^2 \right\} \\ &= \underbrace{E_{\mathcal{D}}\left\{ \left(\widehat{G}_{\mathrm{HO}}(\widehat{\boldsymbol{w}}) - G(\widehat{\boldsymbol{w}}) \right)^2 \right\}}_{\text{variance}} + \underbrace{E_{\mathcal{D}}\left\{ (G(\widehat{\boldsymbol{w}}) - G(\boldsymbol{w}^*))^2 \right\}}_{\text{bias}}. \end{aligned} \tag{5}$$

where $G(\boldsymbol{w}^*)$ is the minimum achievable error within the model, i.e., $\boldsymbol{w}^* = \operatorname{argmin}_{\boldsymbol{w}} G(\boldsymbol{w})$. The bias term is the excess generalization of our model, and decreases with γ. The variance term measures the reliability of the estimator and it increases when γ decreases. We therefore expect an optimal γ to solve the bias/variance trade-off: $\gamma_{\mathrm{opt}} = \operatorname{argmin}_{\gamma} \mathrm{MSE}_{\mathrm{HO}}(\gamma)$. This optimal choice has been studied asymptotically for non-linear models [11], using Vapnik-like bounds [7], and in the context of pattern recognition [6]. Surprisingly, $\gamma_{\mathrm{opt}} \to 1$ as $N \to \infty$, indicating that most data should be used for testing. For finite sample sizes, theoretical investigations are limited to simple models (see below).

***K*-Fold Cross-Validation (KCV).** The average over all training sets in (3) is simulated by resampling the design and test set. In KCV, the data set is split into K disjoint subsets $\mathcal{F}_j$ of approximately equal sizes, $\bigcup_{j=1}^{K} \mathcal{F}_j = \mathcal{D}$. For $\gamma < 1/2$, the split ratio is the ratio of the size of the subsets to the total amount of data, i.e., $K = \lfloor 1/\gamma \rfloor$. We evaluate on each subset the model designed on the remaining data $\mathcal{E}_j = \mathcal{D} \setminus \mathcal{F}_j$.[3] The cross-validation estimator is obtained by averaging the K estimates of generalization error:

$$\widehat{\Gamma}_{\mathrm{KCV}} = \frac{1}{N} \sum_{j=1}^{K} \sum_{k \in \mathcal{F}_j} \ell(\boldsymbol{y}(k), \widehat{\boldsymbol{y}}^{-j}(k)) \tag{6}$$

where $\widehat{\boldsymbol{y}}^{-j}$ is the model designed without subset $\mathcal{F}_j$. It is easy to show that $\widehat{\Gamma}_{\mathrm{KCV}}$ is an unbiased estimate of $\Gamma = E_{\mathcal{E}}\{G(\widehat{\boldsymbol{w}})\}$, the average generalization

[2] For practical reasons γN is restricted to be an integer, i.e., $\gamma = i/N$ where $i = 1, 2, \cdots, N-1$.

[3] For $\gamma > 1/2$ the roles of the design and test set are inverted such that we design on each subset and test on the remaining data.

error based on $N_{\mathcal{E}}$ data. The quality of (6) is assessed by:

$$\begin{aligned}
\mathrm{MSE}_{\mathrm{KCV}}(\gamma) &= E_{\mathcal{D}}\left\{\left(\widehat{\Gamma}_{\mathrm{KCV}} - G(\boldsymbol{w}^*)\right)^2\right\} \\
&= \underbrace{E_{\mathcal{D}}\left\{\left(\widehat{\Gamma}_{\mathrm{KCV}} - \Gamma\right)^2\right\}}_{\text{variance}} + \underbrace{E_{\mathcal{D}}\left\{(\Gamma - G(\boldsymbol{w}^*))^2\right\}}_{\text{bias}}. \qquad (7)
\end{aligned}$$

Randomized permutation cross-validation (PCV). This involves resampling test sets by randomly selecting $N_{\mathcal{F}} = N\gamma$ samples for the test set, and the rest for the design set. This can be repeated at most $K \leq \binom{N}{N_{\mathcal{F}}}$ times. For each permutation, a model $\widehat{\boldsymbol{y}}_j$ with weights $\widehat{\boldsymbol{w}}_j$ is designed, then averaging the K empirical estimates of generalization obtained on each test set yields $\widehat{\Gamma}_{\mathrm{PCV}} = K^{-1}\sum_{j=1}^{K}\widehat{G}(\widehat{\boldsymbol{w}}_j)$. The mean squared error $\mathrm{MSE}_{\mathrm{PCV}}(\gamma)$ is defined as in (7).

Model Selection/Optimization

The design of a model is usually done by estimating model parameters on the training data $\mathcal{T}$, and selecting among alternative models, doing early stopping or tuning various additional hyper parameters on the basis of the validation set $\mathcal{V}$. Either of the 3 methods described above (HO, KCV and PCV) can be used for that purpose. However, the relevant criterion for choosing the optimal split ratio γ should now be the performance of the resulting model. In non-parametric modeling, the ultimate goal is usually to obtain good generalization. The optimal split ratio will then be the value of γ for which the resulting model minimizes the "true" (average) generalization error. In the context of model selection:

$$\gamma_{\mathrm{opt}} = \operatorname*{argmin}_{\gamma} E_{\mathcal{E}}\left\{G_{\widehat{\mathcal{M}}}(\gamma)\right\} \qquad (8)$$

where $G_{\widehat{\mathcal{M}}}(\gamma)$ is the generalization of the model $\widehat{\mathcal{M}}$ which minimizes the cross-validation estimator with split ratio γ. On the other hand, in the context of feature or model selection, the optimal split ratio is one which maximizes the probability of selecting the "correct model". However, as mentioned in e.g., [13, sec. 7.4], selecting a model according to estimated generalization error typically does not result in a consistent selection, i.e., the probability of selecting the correct model does not tend to one. Typically, oversized models will be selected. We shall indeed illustrate in the following examples that those two goals, good generalization and consistent model selection, potentially lead to conflicting decision rules regarding γ.

LOCATION PARAMETER MODEL

In this simple setting, we consider a simple Gaussian variable $y \sim \mathcal{N}(w^\circ, \sigma^2)$ with known σ. This problem has been extensively studied, e.g., [4, 5, 12]. The true generalization of a candidate parameter $\widehat{w}$ is simply $G(\widehat{w}) = \sigma^2 + (w^\circ - \widehat{w})^2$, such that the minimum achievable generalization is σ^2. For the model selection we consider two models: $\mathcal{M}_1$ is a Gaussian with unknown mean estimated from the data, while $\mathcal{M}_0$ is a Gaussian variable with fixed, zero mean $\mathcal{N}(0, \sigma^2)$ (the pruned model).

Design/Test split. For HO cross-validation, tedious but straightforward calculations lead to:

$$\begin{aligned}
\mathrm{MSE}_{\mathrm{HO}}(\gamma) &= \frac{2\sigma^4}{N\gamma}\left(1 + \frac{2}{(1-\gamma)N}\right) + \frac{3\sigma^4}{(1-\gamma)^2 N^2} \\
\gamma_{\mathrm{opt}} &= 1 - \frac{8}{A^{1/3}} + \frac{A^{1/3}}{6N} + \frac{2}{3NA^{1/3}} + \frac{1}{3N} \qquad (9) \\
\text{with } A &= -324N^2 - 144N + 8 + 12N\sqrt{3(243N^2 + 472N - 28)}
\end{aligned}$$

Accordingly, the optimal split ratio converges rather slowly towards 1, as $1 - \gamma_{\mathrm{opt}} = O(N^{-1/3})$, $N \to \infty$. This means, in order to obtain an accurate HO estimate of the generalization error, one should asymptotically reserve the bulk of the data for validation. This is confirmed by the experiments reported in figure 1 (left). All curves are averaged over 40000 replication of the data for each size. When N increases, the optimal γ increases towards 1. Note that the MSE curves flatten, indicating that a wide interval of possible split ratios are near optimal (see also [7]). For K-fold CV:

$$\mathrm{MSE}_{\mathrm{KCV}}(\gamma) = \begin{cases} \dfrac{\sigma^4(2\gamma^3 N - 2\gamma^2 - 6N\gamma^2 + 7\gamma + 6N\gamma - 7 - 2N)}{N^2(\gamma-1)^3} & \gamma \le 0.5 \\ \dfrac{\sigma^4(-4N\gamma^2 - 9\gamma + 8 + 2N\gamma + 2\gamma^2 + 2\gamma^3 N)}{N^2(\gamma-1)^2\gamma} & \gamma \ge 0.5 \end{cases} \qquad (10)$$

It is easy to see that $\partial\mathrm{MSE}_{\mathrm{KCV}}(\gamma)/\partial\gamma > 0$ for all $0 \le \gamma \le 1$ and N. That is, $\mathrm{MSE}_{\mathrm{KCV}}$ is minimum for $\gamma_{\mathrm{opt}} = 1/N$, i.e., leave-one-out (LOO) – irrespective of the size of N. This is supported by the simulation results of figure 1 (right). An interesting feature of these curves is the discontinuity in slope for $\gamma = 1/2$, due to change from overlapping design sets to overlapping test sets. The use of PCV gives estimators with uniformly lower (or equal) MSE compared to KCV (for K sufficiently large). This, however, does not change the qualitative result, as the minimum MSE is always reached for leave-one-out.

Model selection. In the case of model selection, the results are only a function of the normalized variable $\theta \equiv w^\circ\sqrt{N}/\sigma$. We therefore use a single sample size $N = 25$, which gives a good compromise between resolution in

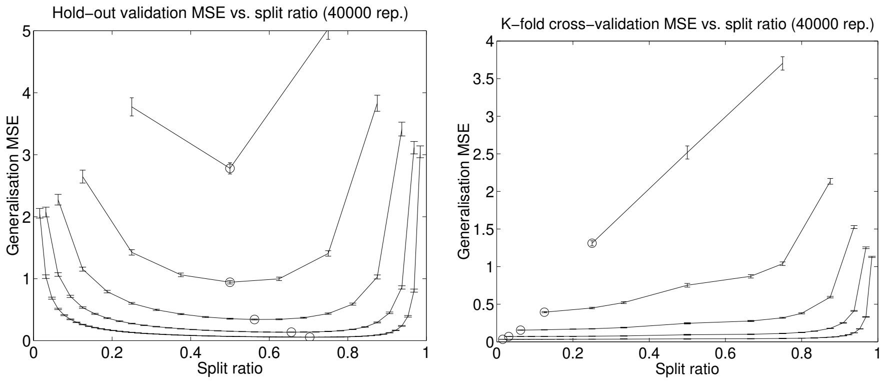

Figure 1: Location parameter example. Left: MSE curves for hold-out cross-validation – the optimal split ratio tends (slowly) to 1 as N increases. Right: K-fold cross-validation – the optimal split ratio is always LOO. In each plot, from top to bottom, $N = 4, 8, 16, 32, 64$ samples. Circle indicates the minimum of each curve, error bars indicate 2 standard deviations.

the possible split ratios and computing requirements, and vary θ through w°. All results are again averaged over 40000 replications of N observations sampled from $\mathcal{N}(w^\circ, 1)$. Figure 2 shows the generalization of the resulting model, for HO, for increasing values of w°. The largest split ratio is optimal for small values of w°. This is illustrated on the right plot by the fact that in that case, the pruned model is almost always selected, and produces a better estimate than the full model, in agreement with [13, sec. 7.4]. For $\theta = 1$, i.e., $w^\circ = \sigma/\sqrt{N} = 0.2$ there is a complete shift and $\gamma = 1/N$ (LOO) becomes optimal. On the left plot, the corresponding curve is almost flat, and at the precise value where the phase transition occurs, both small and large values of γ are within error bars of the optimum. This phase transition occurs because for moderate w° there are two ways of getting good performance: either by selecting the pruned model (for which the excess generalization error is $(w^\circ)^2$), or having enough data for the true model (with average excess generalization error σ^2/N) to perform well. As w° increases, the latter will outperform the former with probability 1. Note that when w° increases some more, the optimal split ratio grows again towards 1 with a very slow asymptotic rate. Figure 2 (right) shows that this is because the split ratio yielding most correct model (consistency) tends to 1. The curves becomes flatter and flatter as w° increases, indicating (as expected) that almost all choices of γ will tend to choose the correct model, and get near-optimal generalization.

The effect of K-fold cross-validation (figure 3) is slightly more subtle. As before, the largest split ratio is optimal for small values of w°. But there are now two transitions, the first one to $\gamma_{\text{opt}} = 1/2$, around $w^\circ = \sigma/\sqrt{N} = 0.2$ and the second one to leave-one-out for a slightly larger value. The first transition occurs as before between the default (pruned) model and the most consistent model. However, due to the overlap in training or validation sets, $\gamma = 1/2$, not LOO, provides the most consistent estimator for small values

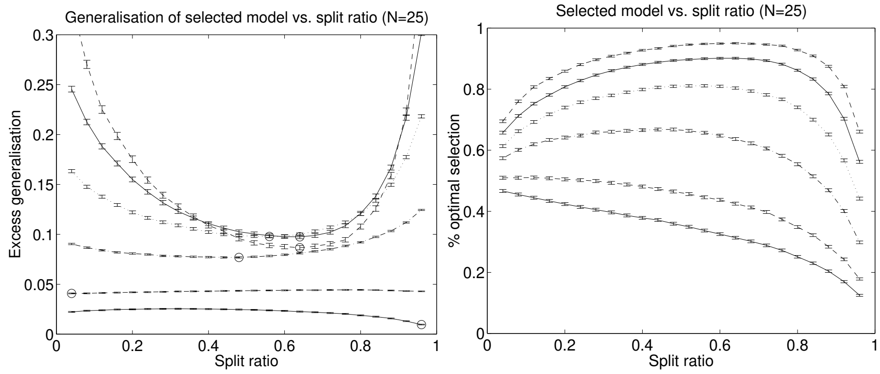

Figure 2: Location parameter example: model selection with hold-out cross-validation, for $w^\circ = 0$ (bottom) to 1 (top) in 0.2 increments. Left: MSE on the hold-out generalization estimator (circle indicates minimum). Right: % correct model selected.

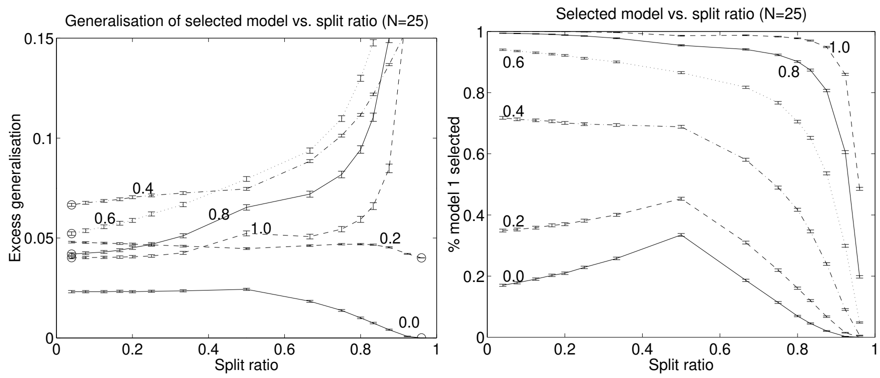

Figure 3: Location parameter example: Model selection with K-fold cross-validation, for $w^\circ = 0$ to 1 in 0.2 increments. Left: excess generalization error of the K-fold generalization estimator (circle indicates minimum). Right: % correct model selected. The curves are labeled with th value of w°.

of w°. The second transition occurs when LOO starts yielding more correct models than 2-fold (fig. 3, right). Additional differences between the K-fold and HO estimators are: 1) in the former, leave-one-out stays optimal as w° grows, 2) the minimum excess generalization error is lower, and 3) the proportion of correctly selected models grows faster towards 1. Note that as the phase transition thresholds are inversely proportional to N, for any $w^\circ \neq 0$, the asymptotically optimal split ratio is $1/N$.

The effect of permutation cross-validation is again similar, and the qualitative conclusion ($\gamma_{\text{opt}} = \text{LOO}$) identical. There is no discontinuity in $\gamma = 1/2$ thanks to the better averaging strategy for intermediate split ratios. There is therefore only one γ-transition, from one extreme value to the other.

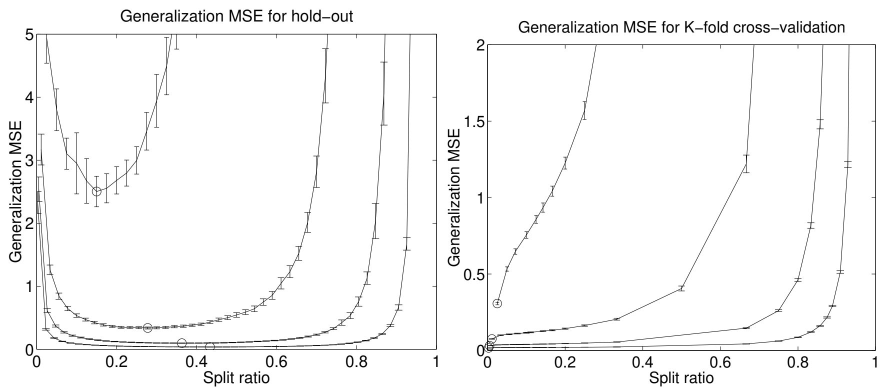

Figure 4: Generalization MSE for the AR filter using hold-out (left) and K-fold (right) cross-validation. From top to bottom, $N = 50, 100, 200, 400$. Circle indicates the minimum of the average MSE, error bars are 2 standard deviations.

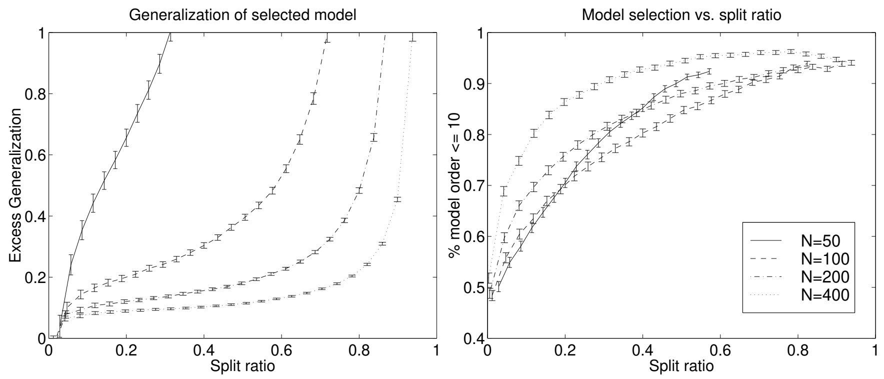

Figure 5: Model order selection. Left: generalization of the selected model. Right: % of non-overestimated models. The legend in the left plot also applies to the right plot.

AUTOREGRESSIVE MODEL

Let us now consider the estimation of a linear autoregressive (AR) filter. The target is a low-pass filter of order 10, with coefficients [-3.99, 8.09, -10.48, 9.42, -6.08, 2.84, -0.94, 0.21, -0.028, 0.0018]. Data are generated by filtering white noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, with $\sigma^2 = 1$. The minimum achievable generalization error $G(\boldsymbol{w}^*)$ is therefore 1.

Design/Test split. In order to assess the optimal split ratio with respect to generalization estimation, we study our cross-validation schemes using a 10th order linear filter. The coefficients are estimated using regularized least squares, $\widehat{\boldsymbol{w}} = \left(\boldsymbol{X}^\top \boldsymbol{X} + \kappa \boldsymbol{I}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$, with $\boldsymbol{X}$ and $\boldsymbol{Y}$ the input and output matrices, respectively, $\boldsymbol{I}$ the unit matrix, and κ set to 10^{-6} times the largest eigenvalue of the covariance matrix. As shown on figure 4, all MSE curves behave in a manner qualitatively similar to the simple Gaussian variable above. Minimum MSE is obtained for increasing γ for HO, and for LOO in the case of K-fold. This suggests the same asymptotics as before: for hold-out $\gamma_{\text{opt}} \to 1$ with $(1-\gamma_{\text{opt}})N \to +\infty$ and $\gamma_{\text{opt}} = 1/N$ for KCV (and PCV). As before, the curves get flatter with increasing N, meaning that a wide range of split ratios become near-optimal.

Model selection. Hold-out cross-validation is used to select the order of the AR model, between 8 and 14. Experiments are reported by averaging over 10000 independent data sets of increasing sizes N = 50, 100, 200 and 400. Figure 5 shows that generalization and model selection lead to conflicting optimal decisions. Clearly, small split ratios give better generalization, but they tend to overestimate the model order. On the other hand, large split-ratios select more parsimonious models but yield poor generalization. Note that due to the small contribution from the last two filter parameters, models of order 8 and 9 are often selected. These results are consistent with [13, sec. 7.4]. This suggests that moderate values of γ might asymptotically provide a good trade-off between model consistency and generalization abilities, though the optimum would depend on a particular weighting of both effects.

NEURAL NETWORKS

We also considered non-linear modeling using neural networks. The target system is the Hénon map: $y(k) = 1-1.4y(k-1)+0.3y(k-2)+\epsilon, \epsilon \sim \mathcal{N}(0,\sigma^2)$ and σ^2 is tuned such that the signal-to-noise ratio is approximately 10%. Data sampled from this system are modeled using a standard feed-forward multi-layer perceptron with one hidden layer of 5 hidden units, and an input layer with time-delayed inputs and task-dependent size.

Experiments indicate that the behaviour of the hold-out estimator is similar to what has been described above. Due to space limitation, the detailed results will be presented at the workshop.

SUMMARY

We addressed the problem of choosing the optimal split ratio for cross-validation estimators. We showed that different cross-validation strategy (the design/test and the training/validation splits), and different objectives (reliable assessment of generalization, best generalization or model consistency) lead to different quality measures (MSE, resulting generalization, probability of correct selection), and potentially result in conflicting decision strategies. For hold-out cross-validation, $\gamma_{\text{opt}} \to 1$ as $N \to \infty$ seems to be well supported theoretically and experimentally. On the other hand, for K-fold and randomized permutation CV, the asymptotically optimal split-ratio is highly dependent on the task and on the model. In particular, we have illustrated that best generalization and model consistency lead to opposite optimal choices

($\gamma_{\text{opt}} \to 0$ and $\gamma_{\text{opt}} \to 1$ respectively).

Acknowledgments. Research supported by the Danish Research Councils through the Danish Computational Neural Network Center (CONNECT), the THOR Center for Neuroinformatics and the European Union, through BIOMED II grant number BMH4-CT97-2775. Lars Kai Hansen is acknowledged for valuable discussions.

REFERENCES

[1] P. Burman, "A Comparative Study of Ordinary Cross-Validation, v-fold Cross-validartion, and the Repeated Learning-Testing Methods," **Biometrika**, vol. 76, no. 3, pp. 503–514, 1989.

[2] B. Efron and R. Tibshirani, **An Introduction to the Bootstrap**, New York, NY: Chapman & Hall, 1993.

[3] S. Geisser, "The Predictive Sample Reuse Method with Applications," **Journal of the American Statistical Association**, vol. 50, pp. 320–328, 1975.

[4] C. Goutte and L. K. Hansen, "Regularization with a pruning prior," **Neural Networks**, vol. 10, no. 6, pp. 1053–1059, 1997.

[5] L. K. Hansen and C. E. Rasmussen, "Pruning from Adaptive Regularization," **Neural Computation**, vol. 6, pp. 1223–1232, 1994.

[6] W. H. Highleynman, "The Design and Analysis of Pattern Reocgnition Experiments," **The Bell Systems Technical Journal**, pp. 723–744, March 1962.

[7] M. Kearns, "A Bound on the Error of Cross Validation Using the Approximation and Estimation Rates, with Consequences for the Training-Test Split," **Neural Computation**, vol. 9, no. 5, pp. 1143–1161, 1997.

[8] M. Kearns, Y. Mansour, A. Y. Ng and D. Ron, "An Experimental and Theoretical Comparison of Model Selection Methods," **Machine Learning**, vol. 27, no. 1, pp. 7–50, April 1997.

[9] J. Larsen, **Design of Neural Network Filters**, Ph.D. thesis, Electronics Institute, The Technical University of Denmark, March 1993.

[10] J. Larsen, L. N. Andersen, M. Hintz-Madsen and L. K. Hansen, "Design of Robust Neural Network Classifiers," in **Proceedings of IEEE ICASSP'98**, IEEE, May 1998, vol. 2, pp. 1205–1208.

[11] J. Larsen and L. K. Hansen, "Empirical generalization assessment of neural network models," in F. Girosi, J. Makhoul, E. Manolakos and E. Wilson (eds.), **Neural Networks for Signal Processing V – Proceedings of the 1995 IEEE Workshop**, Piscataway, New Jersey: IEEE, 1995, pp. 42–51.

[12] J. Larsen, C. Svarer, L. N. Andersen and L. K. Hansen, "Adaptive Regularization in Neural Network Modeling," in **Neural Networks: Tricks of the Trade**, Germany: Springer-Verlag, 1998, no. 1524 in Lecture Notes in Computer Science.

[13] J. Shao and D. Tu, **The Jackknife and Bootstrap**, New York, NY: Springer-Verlag, 1995.

[14] M. Stone, "Cross-validatory choice and assessment of statistical predictions," **Journal of the Royal Statistical Society B**, vol. 36, no. 2, pp. 111–147, 1974, with discussion.