Generalisation

Adrian Bevan

email: a.j.bevan@qmul.ac.uk

Queen Mary
University of London

Generalisation

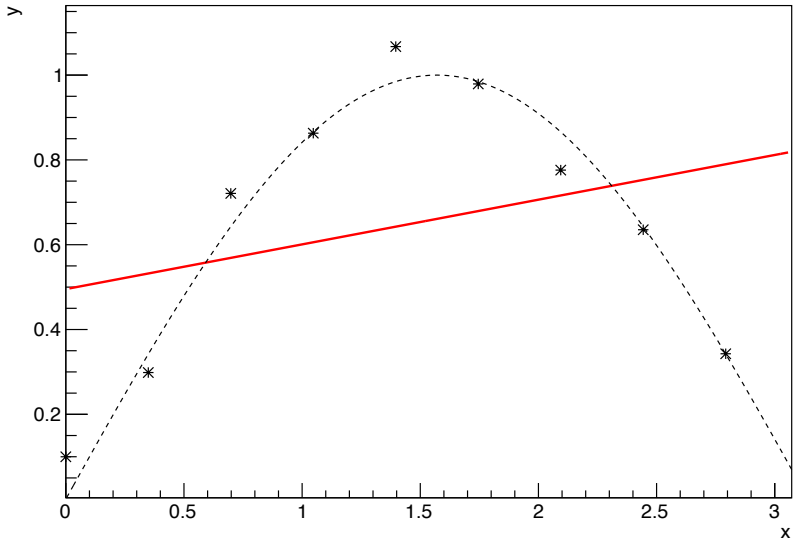
- We generally start from a point with limited statistics to train an MVA.
- How do we know that the MVA trained is sufficiently general to behave well when applied to new data?
- This problem has several possible solutions; here we cover
  - Regularisation
  - Cross validation (CV)
- With regard to selection of MVAs we also discuss the concept:
  - Committee of MVAs

Generalisation

- One can consider curve fitting as an example of over fitting (following the discussion in Bishop, Chapter 9).
  - Given N data points generated with noise randomly from some underlying distribution f(x), we can approximate the function with a polynomial.
  - Increasing the order of the polynomial means that we can ultimately obtain a perfect fitting result at the data points.
  - However this provides us with a bad approximation of the function.

Generalisation

- Example:
  - y=sin(x)
  - noise term = Gaussian with a width of 0.1
  - Sample of 9 points, equally spaced in x



Dashed line corresponds to y=sin(x)

Data are plotted as points (*)

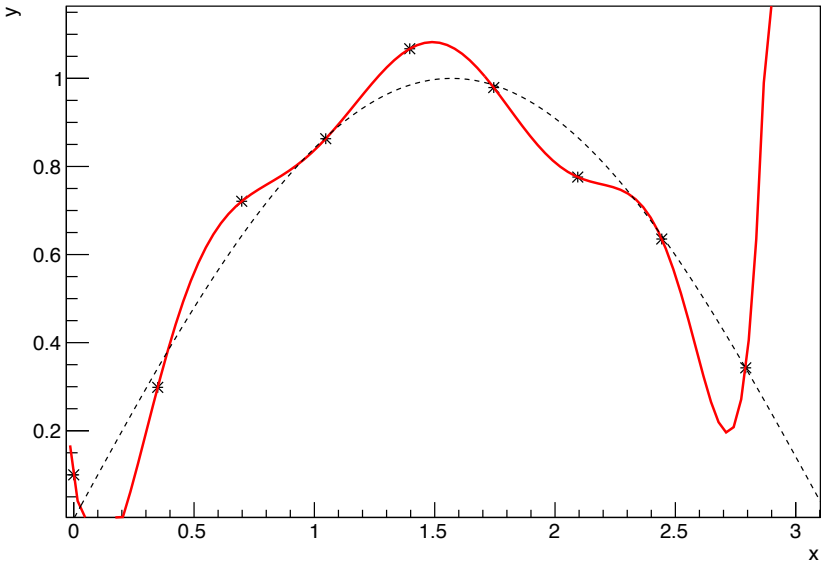
red line is a polynomial fit:

$$\widehat{y} = a + bx$$

There is poor agreement between the fit model and the data as the model is too simplistic. We can increase the order of the polynomial to improve this.

Generalisation

- Example:
  - y=sin(x)
  - noise term = Gaussian with a width of 0.1
  - Sample of 9 points, equally spaced in x



Dashed line corresponds to y=sin(x)

Data are plotted as points (*)

red line is a polynomial fit:

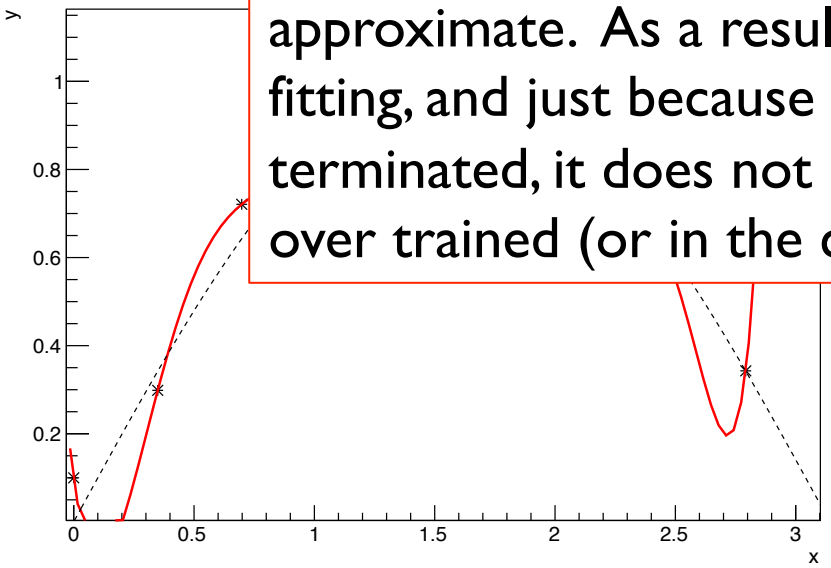
$$\widehat{y} = \sum_{i=0}^{9} a_i x^i$$

There is perfect agreement between the data and model at 9 of the points, but the model is poor between points.

Generalisation

- Example:
  - y=sin(x)
  - noise term = Gaussian with a width of 0.1
  - Samp

An MVA generally has many more dimensions (defined by the number of weight parameters) to approximate. As a result MVAs are prone to over fitting, and just because a training cycle has terminated, it does not mean that the result is not over trained (or in the case of SVMs over fitted)



y=sin(x)

$$\widehat{y} = \sum_{i=0}^{9} a_i x^i$$

There is perfect agreement between the data and model at 9 of the points, but the model is poor between points.

Regularisation

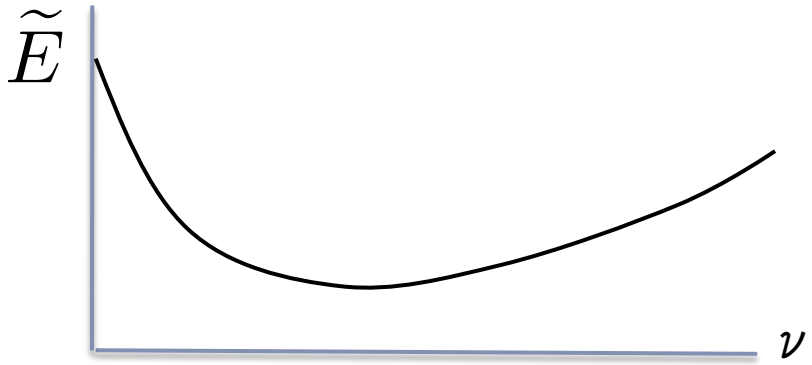
- The problem is that the model description, with respect to the underlying true description, has both a bias and variance (e.g. see Section 9.2 of Bishop).
- Over fitting is what happens when the model is fine tuned to minimise the bias at the cost of variance.
- We need to balance the two competing factors.
- We can perform regularisation to alleviate this issue by adding a penalty term to the training error of our MVA.
- A commonly used simple form of regularisation for neural networks is weight-decay:

$$E \to \widetilde{E} = E + \nu\Omega$$

$$\Omega = \frac{1}{2}\sum_i \omega_i^2$$

Regularisation

- ν is a tuneable regularisation parameter used to help smooth out the weight contributions.
- ν can be optimal, too large or too small and is another factor that needs to be taken into account when training.
- In order to determine the optimal value of ν we have to train the MVA many times, scanning through different values of this regularisation parameter.
  - The optimal value minimises the error



Cross validation (CV)

- Given a limited amount of data to perform a supervised learning training of an MVA, how can we obtain an MVA solution that is sufficiently general when applied to new data, but at the same time makes optimal use of limited statistics.

- Discuss:
  - Hold out CV
  - k-fold CV
  - leave one out CV (leave p-out CV is a trivial extension)

Devroye and Wagner (1979). Distribution-Free performance Bounds for Potential Function Rules. IEEE Transaction in Information Theory, 25 (5) 601–604.

Hold out method

- Divide the data into two samples.
- Use one sample to *train* and one sample to *validate*.

- Logic is that an MVA performing similarly for the training and validation sets will be robust. However it is possible that one obtains an MVA that is fine tuned to the noise found in the validation sample.
  - One can check for this using a third "test" sample of data.

- Also known as simple validation method.
  - This is the most commonly used type of validation for MVAs performed in particle physics. One a few analyses use a more sophisticated validation method*.

* at the time of writing, 2015, based on the past 15 years of common practice in this sub-field of physics.

k-fold CV

- The hold out method does not use the validation data to determine the parameters of the MVA of interest.
  - These are use to compute the error as a function of training evolution to check against that of the training set.
- For situation where known samples of events are scarce, this can be viewed as a wasteful approach.
- The concept of CV was developed to overcome this issue (an provide a more generalised solution for the MVA).
- CV involves averaging the several hold out estimators.

Geisser, S. (1975). The predictive sample reuse method with applications. J. Amer. Statist. Assoc., 70:320–328.

- k-fold CV is just one type of CV method.

k-fold CV

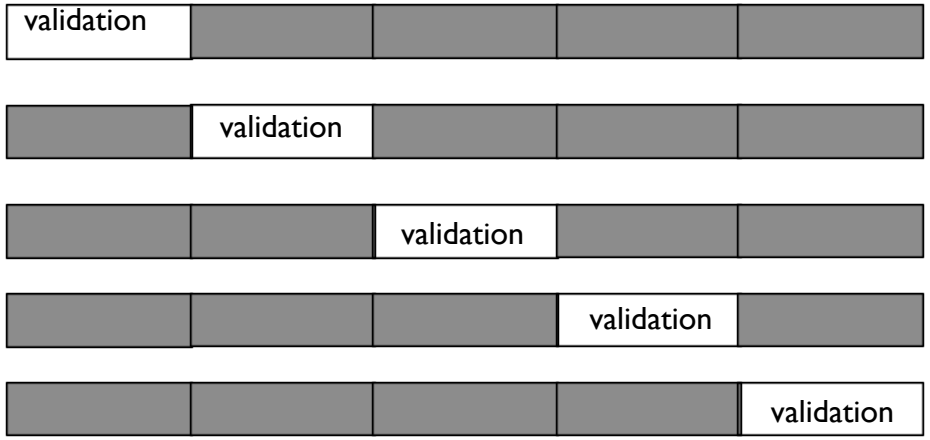
1) Reserve a test sample from the data $\Omega \rightarrow \Omega' \subset \Omega$ (if one wants to validate generalisation beyond the *k*-fold cross validation step).
2) Randomly split the remaining data into k sub samples:
   $\Omega' \rightarrow \Omega_i, i = 1, 2, \ldots k$.
3) Cycle through training *k* times, each time leaving one sub-sample out.

e.g. 5-fold cross validation: train 5 times dropping out one sub-sample at a time.

Use average MVA parameter configuration obtained from the *k*-folds.

The optimal value of k needs to be determined.

limiting case:
k=N(data): gives the leave-one-out cross validation method.



Leave one out CV

- In the extreme limit of k-fold CV that $k \rightarrow N(data)$ we obtain the leave one out CV method.
- Requires N(data) trainings of an MVA.
- Average the result obtained from the N(data) MVAs to determine the output.
- Can provide useful results for small samples of data where training and validation examples are scarce.
- However, can be computationally expensive for large data samples.
- See: Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. J. Roy. Statist. Soc. Ser. B, 36:111–147. With discussion and a reply by the authors; Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. Technometrics, 16:125–127; Geisser, S. (1975). The predictive sample reuse method with applications. J. Amer. Statist. Assoc., 70:320–328.
- This can be extended to the leave p-out CV method, where one successively omits p examples from a training and cycles through the possible permutations.

Shao, J. (1993). Linear model selection by cross-validation. J. Amer. Statist. Assoc., **88** (422):486–494.

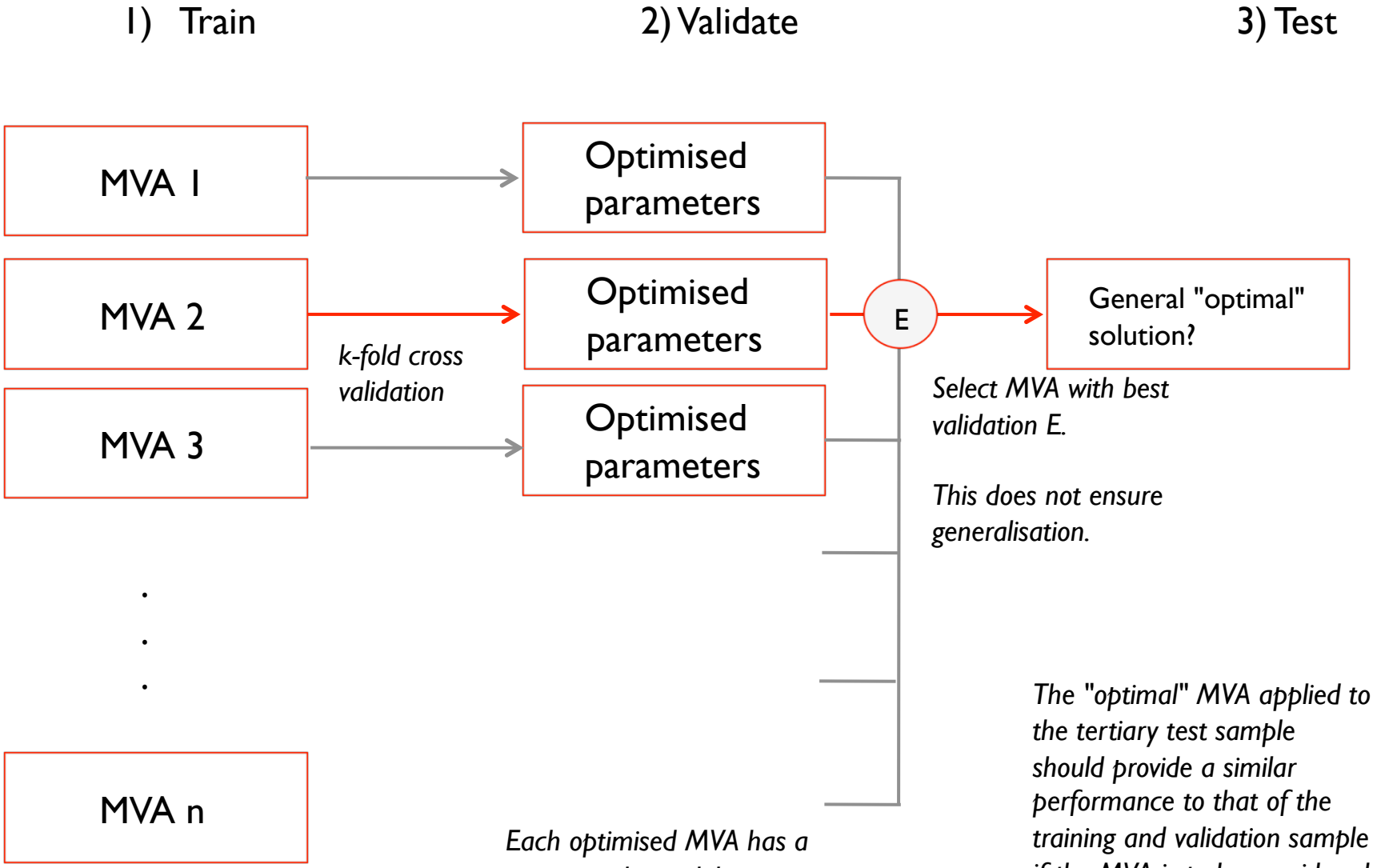
What value of k should you use?

- Possible values of k range from 2 to N(data).
- It is not obvious what value of k will produce the best training of a given MVA.
- Just as the parameters of the MVA merit optimisation via training or fitting, so the value of k also merits optimisation.
- Note for example: the libsvm SVM implementation in R uses a 10-fold CV by default.
- A study, varying k can be found in:

Kohavi (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, Proc. 14th Int. Joint Conf. Artif. Intell. Vol 2, Morgan Kaufmann.

How can we avoid being fine tuned to the validation set?

- If there is still a concern that he MVA training/fitting may be susceptible to over training or over fitting on the validation sample; one can protect against this by reserving a portion of data as an independent test sample.
- The training cycle becomes:
  1. Divide the data into training validation and test samples.
  2. Use the training and validation samples to develop the MVA and categorise the "best" one.
  3. Use the test sample performance to check for fine tuning of the MVA as a final step.
     - If the chosen MVA is over trained / over fitted, then go back to the start.


1) Train
2) Validate
3) Test
MVA 1
MVA 2
MVA 3
MVA n
k-fold cross validation
Optimised parameters
Optimised parameters
Optimised parameters
E
General "optimal" solution?
Select MVA with best validation E.
This does not ensure generalisation.
MVA algorithms with different generic architectures, and weight parameters to be determined.
Each optimised MVA has a corresponding validation error E. This FOM can be used to determine the "optimal" MVA to use with the test sample.
The "optimal" MVA applied to the tertiary test sample should provide a similar performance to that of the training and validation sample if the MVA is to be considered general.
If not; start the process again.

Committee

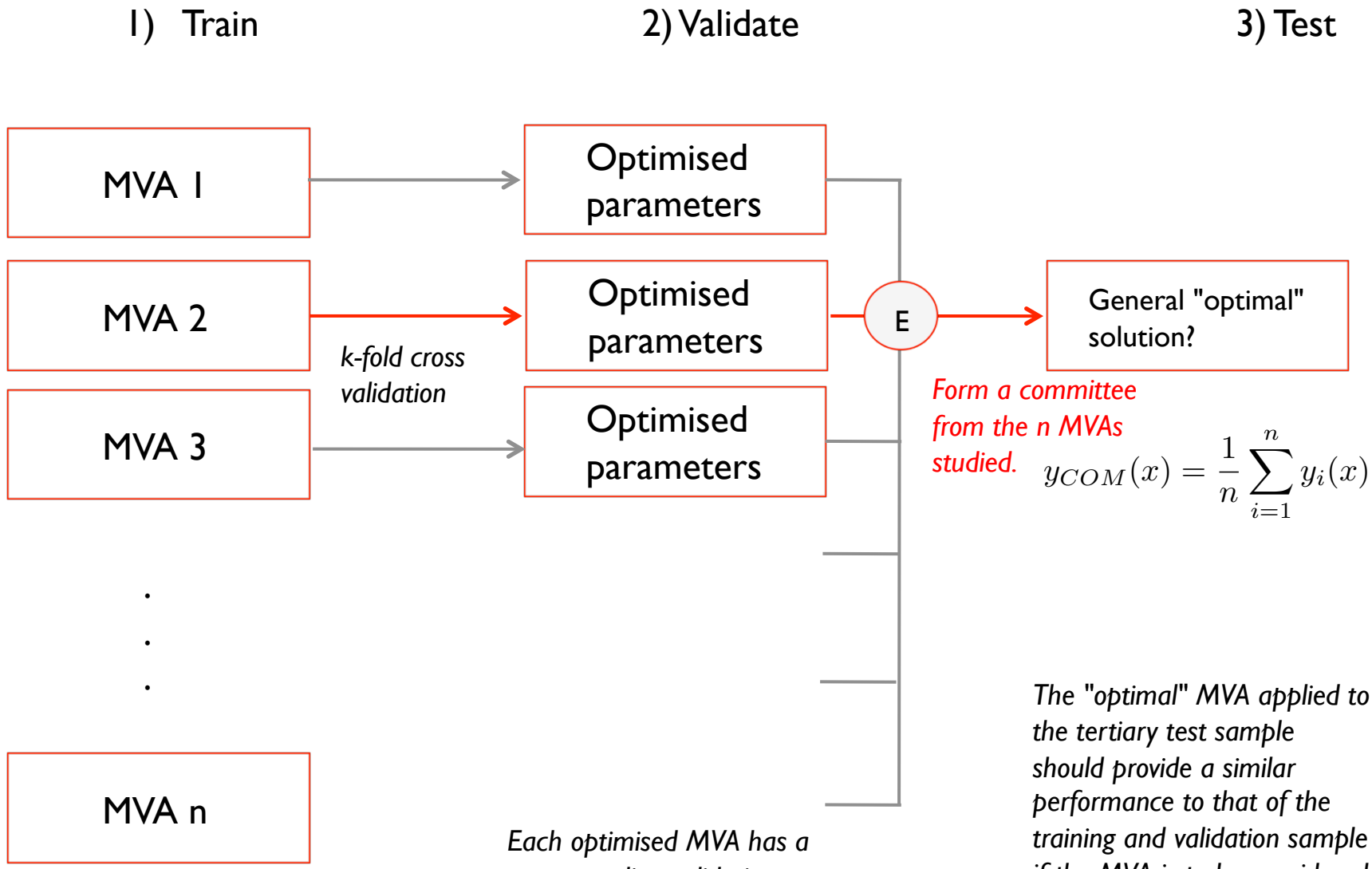
- It is a waste of effort to train n different MVAs and just select one of them to use to distinguish between signal and background.

- Instead one can average the results

$$y_{COM}(x) = \frac{1}{n}\sum_{i=1}^{n} y_i(x)$$

Perrone and Cooper (1993), When networks disagree: ensemble methods for hybrid neural networks. Artificial NN for speech and vision pp 126-142, Chapman and Hall; Perrone (1994) General averaging results from convex optimisation in Mozer et al (eds) Procs 1993 Connectionist Models Summer School 364-371.

- Can in general provide a better separation between signal and background than an individual MVA.
- Neglects the fact that different MVAs will have different error rates; but one can weight contributions accordingly.


1) Train
2) Validate
3) Test
MVA 1
MVA 2
MVA 3
MVA n
k-fold cross validation
Optimised parameters
Optimised parameters
Optimised parameters
E
General "optimal" solution?
Form a committee from the n MVAs studied.
$y_{COM}(x) = \frac{1}{n}\sum_{i=1}^{n} y_i(x)$
MVA algorithms with different generic architectures, and weight parameters to be determined.
Each optimised MVA has a corresponding validation error E. This FOM can be used to determine the "optimal" MVA to use with the test sample.
The "optimal" MVA applied to the tertiary test sample should provide a similar performance to that of the training and validation sample if the MVA is to be considered general.
If not; start the process again.

Summary

- Regularisation and cross validation provide methods that can be applied to MVA training in order to reduce the issue of over training or over fitting.

- These lead to further complication in the process of setting up and defining the MVA of interest, but also lead to new ideas on how to select or combine information from MVAs such as the committee concept.

- The issue of generalisation is something that is largely neglected in some fields where MVA methods are applied to data and are far removed from machine learning;
  - that's not necessarily a bad thing as long as at least some elementary procedure such as the hold out method is employed in an attempt to avoid over training.
  - Using cross validation or regularisation should lead to a more robust result, and in turn may yield better performance over the simple holdout method.

Suggested further reading

- The references indicated throughout these slides.
- Bishop, Neural Networks for Pattern Recognition (1995) OUP, Chapter 9.
- Porter and Narsky, Statistical Analysis Techniques in Particle Physics (2013) Wiley (various sections – see index).
- CV:
  - See the recent review by S. Arlot and A. Celisse on "Cross-validation procedures for model selection" in Statistics Surveys Vol. **4** (2010) 40–79, and references therein for a detailed discussion on CV.