AKI VEHTARI *(Helsinki University of Technology, Finland)*

My comments concern the use of cross-validation for model comparison. Authors have used 5-times-repeated 10-fold-CV (Algorithm 1) for model comparison, but they do not mention how are the results in Figure 1 used to tell whether the difference between two models is significant. Five repetitions in Algorithm 1 produce samples from the distribution in which the variability comes from the internal randomness in the algorithm (including such things as how sensitive the approach is to initial values and properties of the stochastic search algorithm) and the variability due to different random divisions in the 10-fold-CV. If the goal is to estimate which model would make best predictions for new data, it is important also to take into account the uncertainty from not knowing the distribution of the future data. Vehtari and Lampinen (2002) describe how to obtain samples from the distribution of the cross-validation estimated expected utility (e.g., deviance) estimate taking properly into account the uncertainty from the internal randomness in the algorithm, the variability due to different random divisions in the k-fold-CV *and* the approximation of the future data distribution. Vehtari and Lampinen (2002) discuss and demonstrate that the uncertainty from the approximation of the future data distribution dominates and the other uncertainties are small (at least for stable models and algorithms). Using the obtained samples from the distributions of the expected utilities, models can be compared in Bayesian way by computing the probability of one model having a better expected utility than some other model (Vehtari & Lampinen 2002). Also when using k-fold-CV it is useful to use correction term. Since for each fold $1/k$ of the data is left out, the expected utility is estimated conditioning only on $1 - 1/k$ of the data and thus uncorrected k-fold-CV provides biased estimate of expected utility conditioned on the full data. This can be corrected using less well known first order correction proposed by Burman (1989) and demonstrated for Bayesian models by Vehtari and Lampinen (2002). This correction is important especially in model assessment but also in model comparison if the models compared have different steepness of the learning curves.

ADDITIONAL REFERENCES IN THE DISCUSSION

Burman, P. (1989). A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods. *Biometrika* **76**, 503–514.

Vehtari, A. and Lampinen, J. (2002). Bayesian Model Assessment and Comparison Using Cross-Validation Predictive Densities. *Neural Computation*, (in press).