Introduction
In prediction and decision problems, it is natural to assess the predictive ability of the model by estimating the expected utilities, that is, the relative values of consequences of using the model (Godd, 1952; Bernardo & Smith, 1994).

Expected Utilities
The posterior predictive distribution of output $y^{(i)}$ for the new input $x^{(i)}$ is given by the training data $D = \{x^{(1)}, y^{(1)}; \ldots, x^{(n)}, y^{(n)}\}$ is obtained by

$$p(y^{(i)^*}|x^{(i)}, D, M) = \int p(y^{(i)^*}|x^{(i)}, \tilde{\theta}, D, M)p(\tilde{\theta}|D, M)\, d\tilde{\theta}. \quad (1)$$

Given a utility function $u$, the expected utility is obtained by taking the expectation

$$\hat{u} = E_{y^{(i)^*}|x^{(i)}, D, M}[u(y^{(i)^*}, x^{(i)}, D, M)]. \quad (2)$$

The expectation could also be replaced by some other summary quantity, like the $\nu$-quantile. Examples of generic utilities are the absolute error (3) and the predictive likelihood (4)

$$u = \text{abs}(E_{y^{(i)^*}|x^{(i)}, D, M}[y^{(i)^*}]) \quad (3)$$

$$u = p(y^{(i)^*}|x^{(i)}, D, M). \quad (4)$$

Cross-Validation Predictive Densities
Expected utilities can be estimated using cross-validation (CV) predictive densities. $k$-fold CV is a robust way of obtaining CV predictive densities for complex hierarchical Bayesian models. As the distribution of $y^{(i)^*}$ is unknown, we approximate it by using the samples we already have. The predictive distribution is replaced with a collection of leave-one-out cross-validation (LOO-CV) predictive densities

$$p(y^{(i)^*}|x^{(i)}, D(i), M), i = 1, 2, \ldots, n, \quad (5)$$

where $D(i)^{-}$ denotes all the elements of $D$ except $x^{(i)}$. To get the expected utility estimate, these predictive densities are compared to the actual $y^{(i)}$'s using the utility $u$, and the expectation is taken over $i$

$$\hat{u}_{\text{LOO}} = E_{D(i), y^{(i)}}[u(y^{(i)^*}, x^{(i)}, D(i), M)]. \quad (6)$$

Importance-sampling leave-one-out cross-validation (IS-LOO-CV)
In IS-LOO-CV, instead of sampling directly from $p(\hat{\theta}|D, M)$, samples $\tilde{\theta}$ from the full posterior $p(\hat{\theta}|D, M)$ are reused (Gelfand, Dey & Chang, 1992; Gelfand, 1996). Compared to sampling from the full posterior distribution the additional computation time is negligible, but importance sampling may be numerically unstable.

$k$-fold cross-validation ($k$-fold-CV)
In $k$-fold-CV, we sample only from $i$ (e.g., $i = 10$) $k$-fold CV distributions $p(\hat{\theta}|D(i), M)$ and get a collection of $k$-fold CV predictive densities

$$p(y^{(i)^*}|x^{(i)}, D(i), M), i = 1, 2, \ldots, n, \quad (7)$$

where $i$ is a set of data points as follows: the data is divided into $k$ groups so that the sizes are as nearly equal as possible and $x^{(i)}$ is the set of data points in group where the $i$th data point belongs. With appropriate grouping $k$-fold CV can be also used when there are finite range dependencies in the data. Since the $k$-fold CV predictive densities are based on smaller training data sets than the full data set, we have used a first order bias correction (Burman, 1989) when computing the expected utilities.

Example 1
Figures 1 and 2 present some results comparing IS-LOO-CV and $k$-fold-CV in a real world problem of predicting the quality properties of concrete. In the study, we had 27 explanatory variables and 215 samples. We tested 10-hidden-unit MLP networks and Gaussian Process (GP) models.

Figure 1: Comparison of IS-LOO-CV and $k$-fold-CV with and without bias correction. IS-LOO-CV fails as importance sampling does not work well in this problem (see Figure 2). $k$-fold-CV without bias correction gives too pessimistic estimates.

Figure 2: IS-LOO-CV fails as the importance weights have large variance. The quality of importance weights can be summarized by estimating the effective sample size $m_{\text{eff}} = 1/\sum w_i^2$ (Kong et al. 1994).

Other Predictive Densities
Relations of CV-predictive densities to other predictive densities can be illustrated by examining the following equations (for simplicity, the covariate $x$ has been dropped out).

- Cross-Validation (expected utility): $\sum \log p(y^{(i)^*}|D(i), M) = E_{\tilde{\theta}}[\log p(y^{(i)^*}|D(i), M)]$
- Marginal posterior (training error): $\sum \log p(y^{(i)^*}|D(i), M)$
- Posterior (posterior Bayes factor): $\sum \log p(y^{(i)^*}|D(i), M) - \log p(D|M)$
- Prior (Bayes factor): $\sum \log p(y^{(i)^*}|D(i), M)$

Distribution of the Expected Utility Estimate
Instead of just making a point estimate, it is important to obtain the distribution of the expected utility estimate, as it describes the uncertainty in the estimate. These distributions can also be used to compare models.

We propose to use Bayesian bootstrap (BB; Rubin, 1981) to obtain samples from the distribution of the expected utility estimate. In BB it is assumed that the posterior probabilities for the samples $z_i$ of a random variable $Z$ have Dirichlet distribution and values of $Z$ that are not observed have zero posterior probability. Sampling from the Dirichlet distribution gives BB samples from the distribution of the distribution of $Z$ and thus samples of any parameter of this distribution can be obtained. We first sample from the distributions of each $z_i$ (variability due to Monte Carlo integration) and then from the distribution of the $z$ (variability due to the approximation of the future data distribution). From the obtained samples, it is easy to compute, for example, credible intervals (CI), highest probability density intervals, and kernel density estimates. The approach can handle arbitrary summary quantities and gives a good approximation also in non-Gaussian cases.

Example 2
Figure 4 shows an example of the distribution of the expected utility estimate in a forest scene classification problem, in which the task is to classify pixels to tree or background. Figure 4 is an example image. We had 48 images, from which we extracted 84 Gabor and statistical features and used a 20-hidden-unit MLP network (Model 1).

Figure 3: An example of typical Finnish forest with pine trees.

Figure 4: The distribution of the expected utility estimate for Model 1.

Model Comparison
The distributions of the expected utility estimates can be used to compare models, for example, by plotting the distribution of $\Delta \hat{u}_{\text{LOO}}$ or computing the probability $p(\Delta \hat{u}_{\text{LOO}}>0)$. Following the simplicity postulate (Jeffreys, 1961), it is useful to start from simpler models and then test if more complex model would give significantly better predictions.

Example 3
Continuing Example 2, Figures 5 and 6 show results comparing Model 1 to Model 2, which had only 18 features selected from the full set.

Figure 5: The distributions of expected utility estimates for models 1 and 2.

Figure 6: The distribution of the difference of expected utility estimates. $p(\Delta \hat{u}_{\text{LOO}}>0) = 0.86$.

Information Criteria and Effective Number of Parameters
$\text{CV}$-predictive approach is more robust and reliable than information criteria.

Information criteria such as AIC, NIC, DIC (Akaike, 1973; Murata et al., 1994; Spiegelhalter et al., 2002) estimate expected utilities using asymptotic approximations and plug-in predictive distributions (maximum likelihood, maximum a posteriori or posterior mean). In the case of information criteria, the distribution of the estimate is not so easy to estimate and thus usually point estimates are used, which leads to selection of unnecessarily large models. An essential part of the modern information criteria (like DIC) is the estimation of the effective number of parameters $p_W$ in the model. In the CV approach, estimate of the $p_W$ is not needed, but can be estimated by the difference of the marginal posterior predictive likelihood and the expected predictive likelihood.

Example 4
Continuing Example 1, Figures 7 and 8 show results comparing $k$-fold-CV and DIC in estimating the expected utilities and the effective number of parameters for four different noise models (Normal ($\mathcal{N}$), Student’s $t$, input dependent Normal (in.dep.$\mathcal{N}$) and input dependent $t$ tested in the concrete quality estimation problem.

Figure 7: Comparison of expected utility estimates with $k$-fold-CV and DIC.

Figure 8: Comparison of the effective number of parameters with $k$-fold-CV and DIC.