Cross-Validation, Information Criteria, Expected Utilities and the Effective Number of Parameters

Aki Vehtari and Jouko Lampinen

HELSINKI UNIVERSITY OF TECHNOLOGY
Laboratory of Computational Engineering

Introduction

- Expected utility
  - estimates the predictive performance of the model
  - possible to use application specific utilities
  - useful in both model assessment and comparison
- Estimation of the expected utility
  - cross-validation
  - information criteria, DIC

Expected utility

- Given
  - the training data $D = \{(x^{(i)}, y^{(i)}); i = 1, 2, \ldots, n\}$
  - a model M
  - a future input $x^{(n+1)}$
  - posterior predictive distribution $p(y^{(n+1)}|x^{(n+1)}, D, M)$

  a utility function u compares the predictive distribution to a future observation

- Examples of generic utilities
  - predictive likelihood $u = p(y^{(n+1)}|x^{(n+1)}, D, M)$
  - absolute error $u = \text{abs}\left(\hat{y}^{(n+1)} - y^{(n+1)}\right)$

- The expected utility is obtained by taking the expectation

$$\bar{u} = E_{(x^{(n+1)}, y^{(n+1)})}\left[u(y^{(n+1)}, x^{(n+1)}, D, M)\right]$$

Estimating expected utility

- The expected utility is obtained by taking the expectation

$$\bar{u} = E_{(x^{(n+1)}, y^{(n+1)})} \left[u(y^{(n+1)}, x^{(n+1)}, D, M) \right]$$

- The distribution of $(x^{(n+1)}, y^{(n+1)})$ is unknown
- Expected utility can be approximated
  - sample re-use $\rightarrow$ cross-validation
  - asymptotic approximations $\rightarrow$ information criteria

Cross-validation

- The expected utility

$$\bar{u} = E_{(x^{(n+1)}, y^{(n+1)})} \left[u(y^{(n+1)}, x^{(n+1)}, D, M) \right]$$

- The distribution of $(x^{(n+1)}, y^{(n+1)})$ is estimated using $(x^{(i)}, y^{(i)})$ and the predictive distribution is replaced with a collection of CV predictive distributions

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

where $D^{(\setminus i)}$ denotes all the elements of D except $(x^{(i)}, y^{(i)})$

- CV predictive distributions are compared to the actual $y^{(i)}$'s using the utility u, and the expectation is taken over i

$$\bar{u}_{\text{CV}} = E_i \left[u(y^{(i)}, x^{(i)}, D^{(\setminus i)}, M) \right]$$

Information criteria

- The expected utility $\bar{u} = E_{(x^{(n+1)}, y^{(n+1)})}\left[u(y^{(n+1)}, x^{(n+1)}, D, M)\right]$
- The predictive distribution is replaced with a “plug-in” predictive distribution

$$p(y^{(n+1)}|x^{(n+1)}, \tilde{\theta}, D, M)$$

- Using second order Taylor approximation we obtain

$$\bar{u}_{\text{NIC}} = E_i\left[u(y^{(i)}, x^{(i)}, \tilde{\theta}, D, M)\right] + \text{tr}(K J^{-1})$$

$K = \text{Var}[\bar{u}(\tilde{\theta})']$, and $J = \text{E}[\bar{u}(\tilde{\theta})'']$. The $\bar{u}(\tilde{\theta})'$ and $\bar{u}(\tilde{\theta})''$ represent the first and second derivatives with respect to θ.

- DIC Makes Monte Carlo approximation

$$2\left(E_\theta[\bar{u}(\theta)] - \bar{u}(E_\theta[\theta])\right) \approx \text{tr}(K J^{-1})$$

$$\bar{u}_{\text{DIC}} = \bar{u}(E_\theta[\theta]) + 2\left(E_\theta[\bar{u}(\theta)] - \bar{u}(E_\theta[\theta])\right)$$

The effective number of parameters

- Using log-likelihood utility multiplied by n
  $L(\tilde{\theta}) = \sum_i \log p(y^{(i)}|x^{(i)}, \tilde{\theta}, D, M)$
  - $\mathrm{tr}(K J^{-1}) = p_{\mathrm{eff}}$, the effective number of parameters
  - $0 < p_{\mathrm{eff}} \leq p$
- p_{eff} is influenced by
  - the amount of the prior influence
  - dependence between the parameters
  - number of the training samples $(p_{\mathrm{eff}} \leq n)$
  - distribution of the noise in the samples
  - the complexity of the underlying phenomenon to be modeled

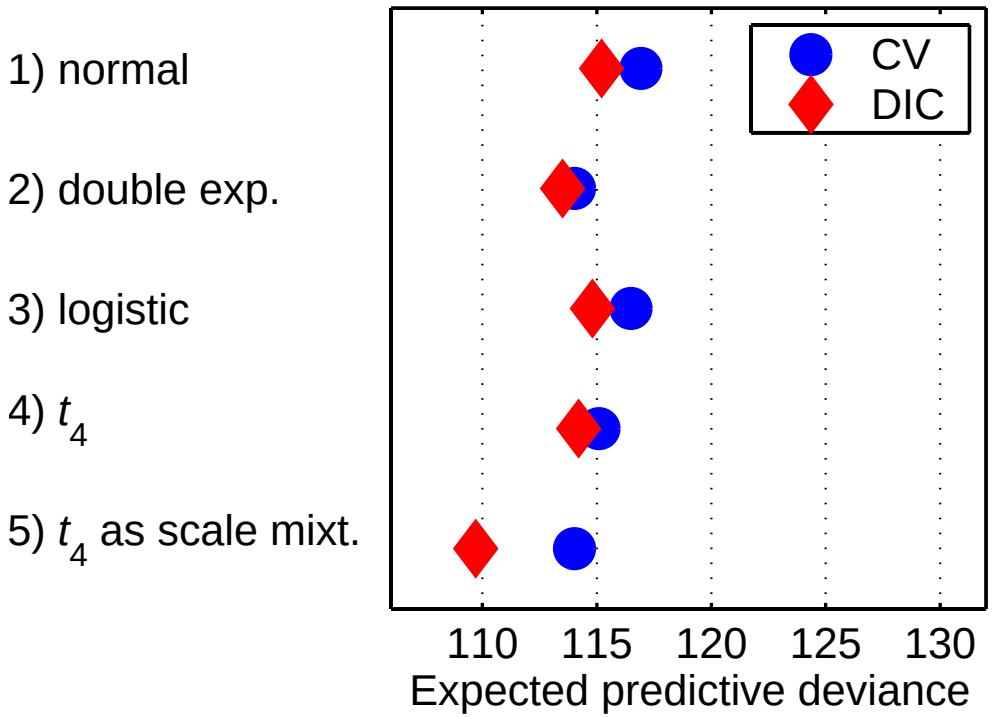
The effective number of parameters

- There is no need to estimate p_{eff} in cross-validation approach
- Using log-likelihood utility multiplied by n

$$
\begin{aligned}
p_{\text{eff,CV}} &= \sum_i \left[\log p(y^{(i)} | x^{(i)}, D, M) \right] - \sum_i \left[\log p(y^{(i)} | x^{(i)}, D^{(\setminus i)}, M) \right] \\
&= L_{\text{MPO}} - L_{\text{CV}}
\end{aligned}
$$

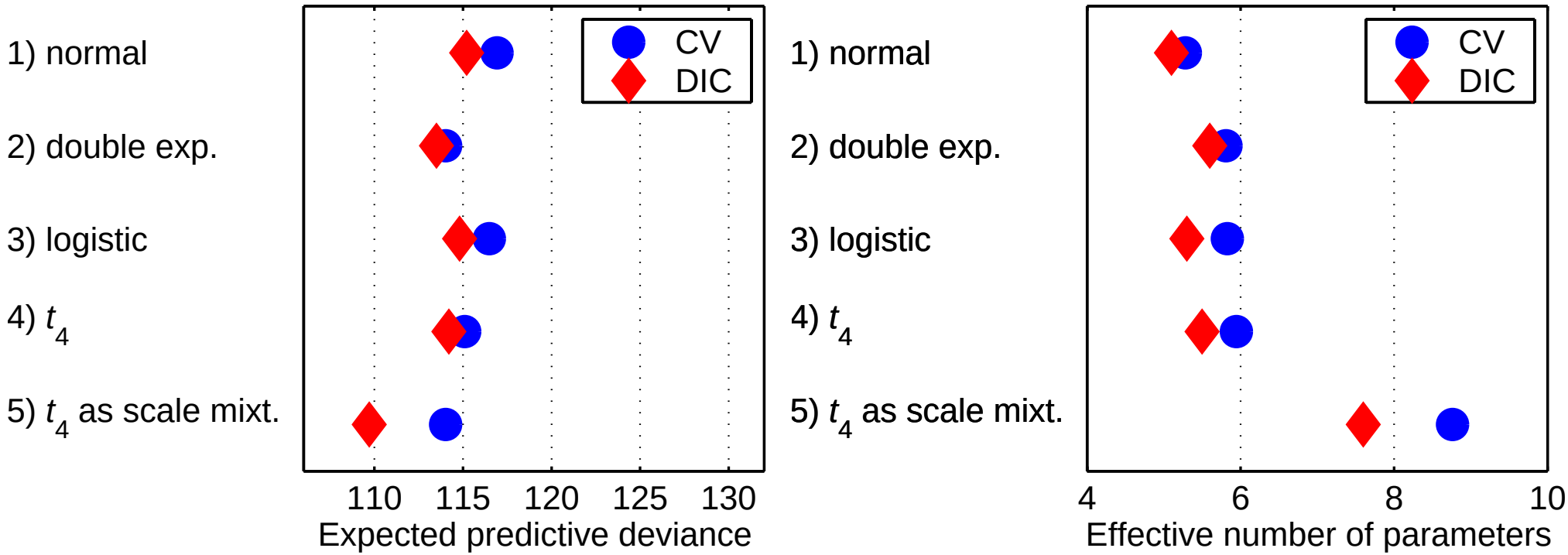
Example

- Robust regression using the stack loss data
  - 3 predictor variables, 21 cases
  - linear regression with 5 different error distribution models



Example

- Robust regression using the stack loss data
  - DIC slightly underestimates the effective number of parameters
  - DIC slightly underestimates the expected predictive deviance



Example

- Robust regression using the stack loss data
  - DIC gives just point estimates
  - In CV approach it is easy to estimate uncertainty

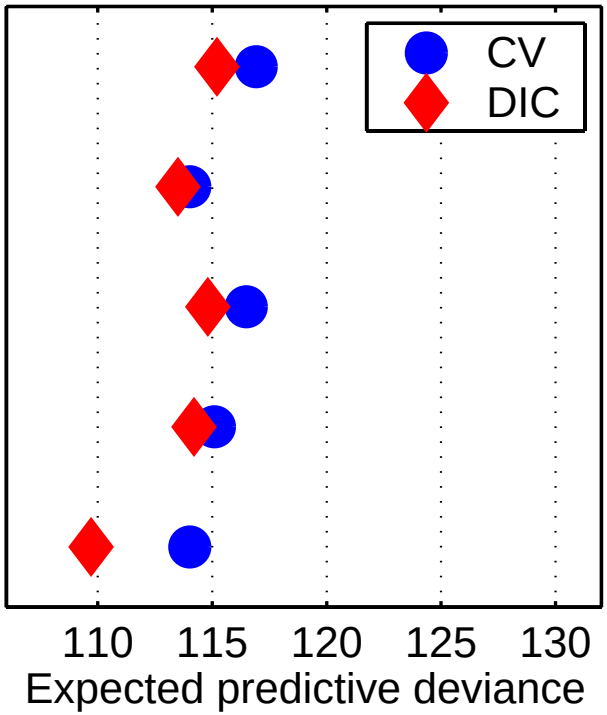
1) normal

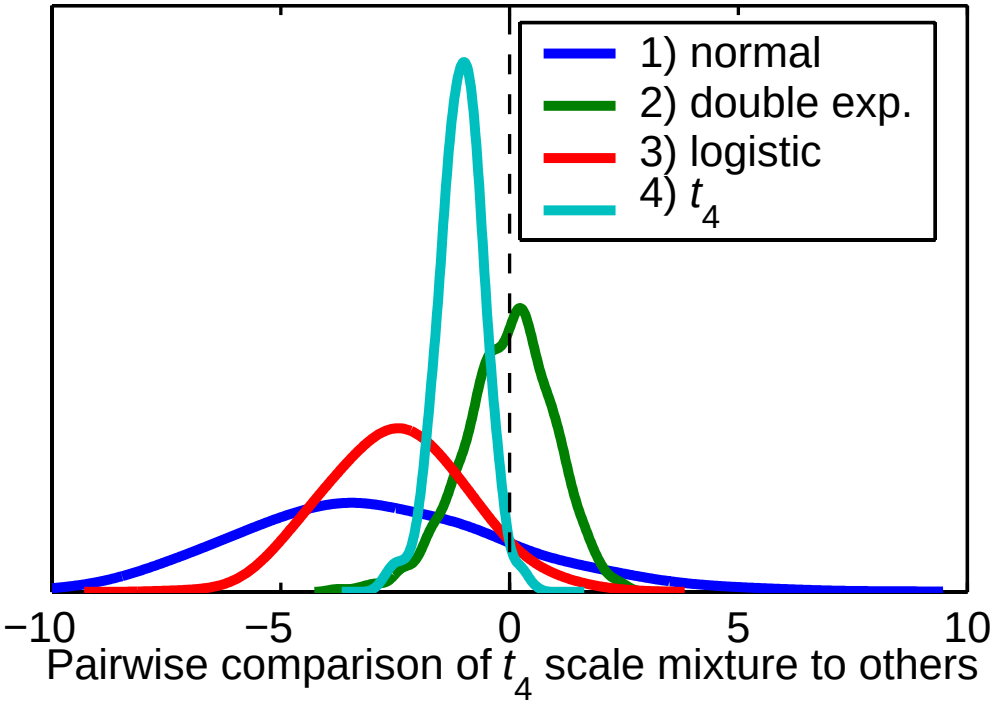
2) double exp.

3) logistic

4) t_4

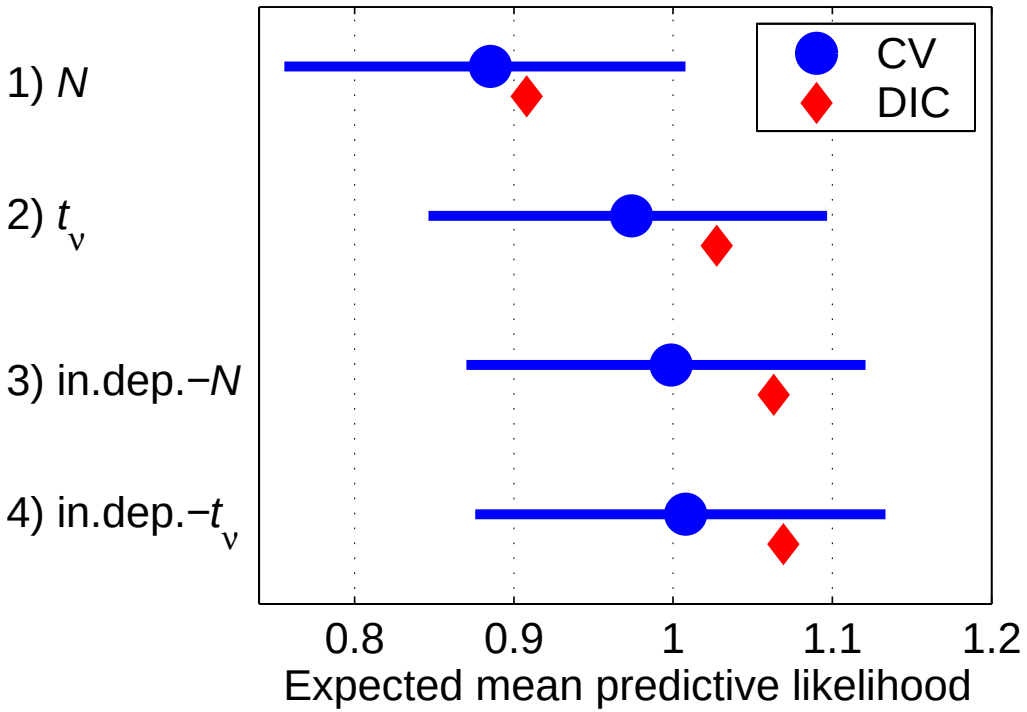
5) t_4 as scale mixt.

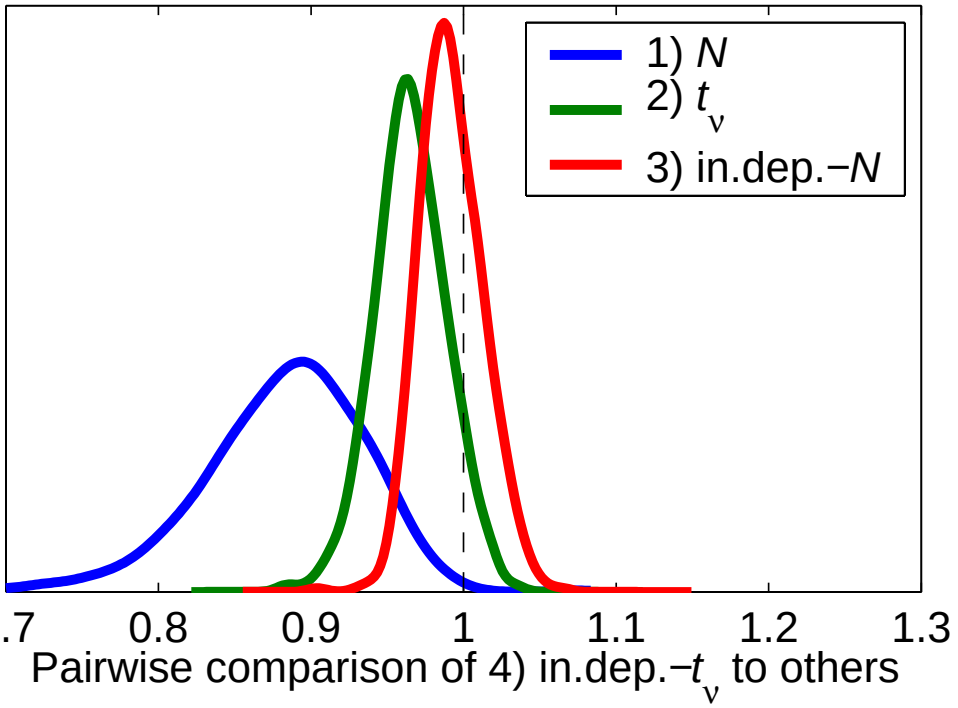




Example

- Concrete quality prediction
  - 27 predictor variables, 215 cases
  - Gaussian process model with 4 different error distribution models



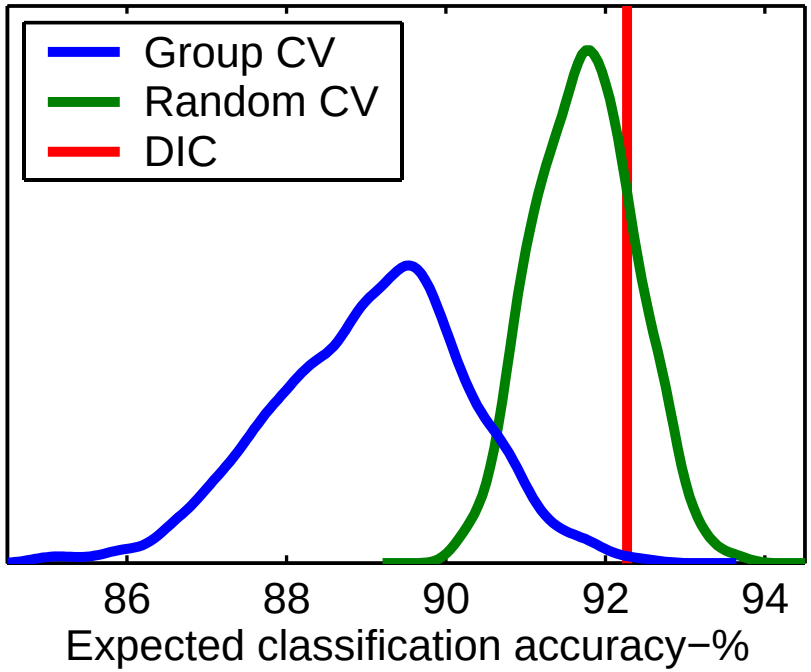


Dependent data

- Different dependencies
  - Group dependencies
  - Time series
  - Spatial
- DIC assumes independence
- CV can handle some finite range dependencies

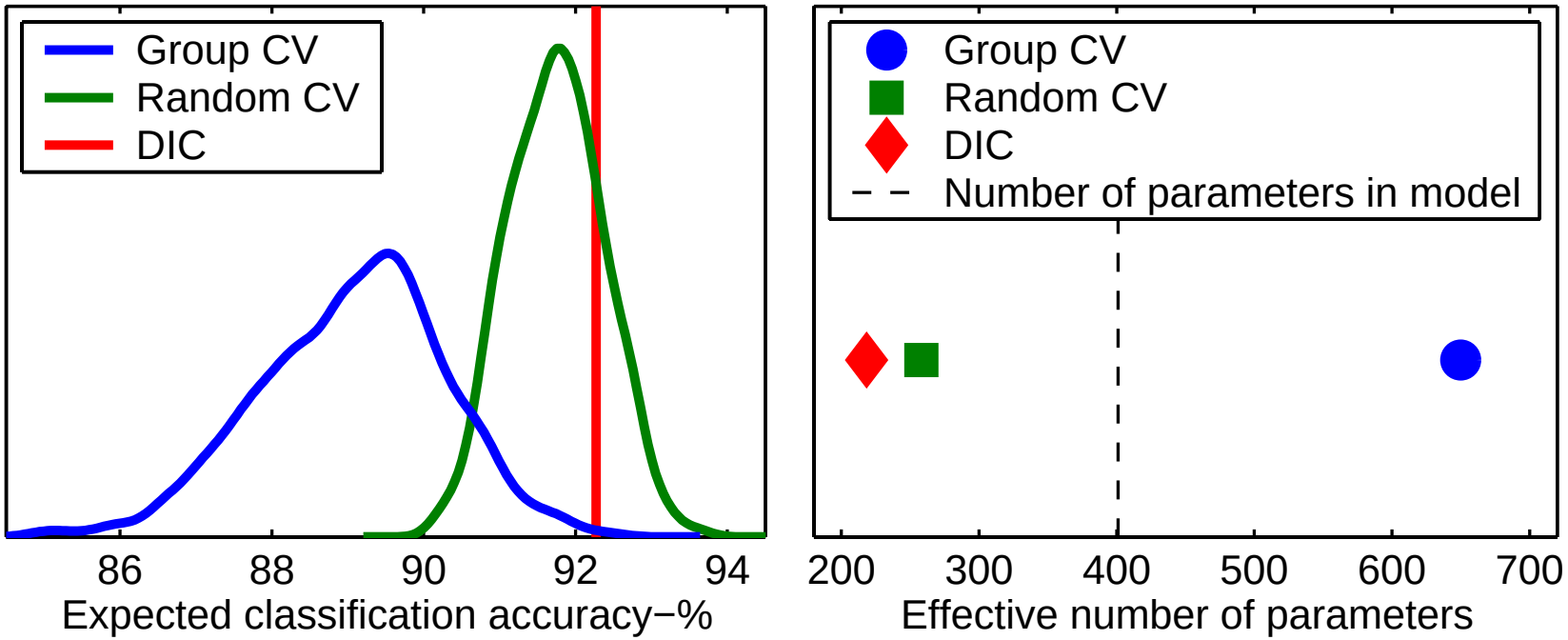
Example of group dependencies

- Forest scene classification
  - 18 predictor variables 48x100 cases
  - 20-hidden-unit MLP with the logistic likelihood model



Example

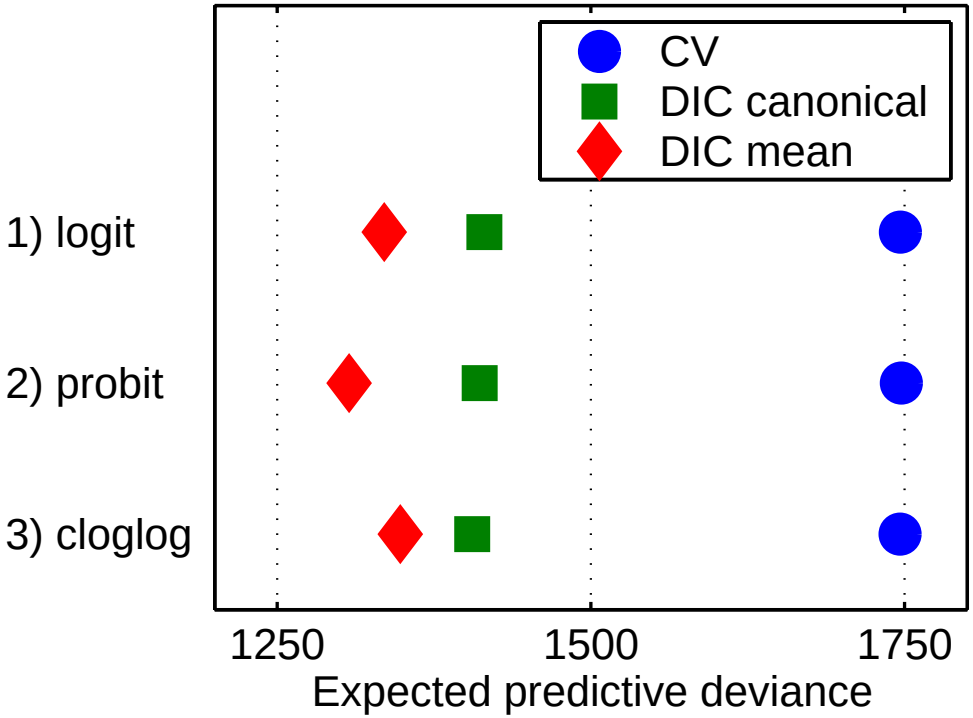
- Forest scene classification
  - DIC assumes independent data points
  - CV can handle group dependencies



Example

- Longitudinal data: the six cities-study
  - 2 predictor variables and 537 children
  - linear model with interaction term, three different link functions and Bernoulli likelihood

	Wheezing status at age of				*Mother smoking*
	7	8	9	10	
Child 1	0	0	0	0	0
Child 2	0	0	1	0	0
Child 3	0	1	0	1	1
⋮					
Child n	1	1	1	1	1



Cross-validation vs. Information criteria

Cross-validation	DIC
• uses full predictive distributions	• uses “plug-in” predictive distributions
• deals directly with predictive distributions	• parametrization problems
• easy to estimate the uncertainty	• estimation of the uncertainty under investigation
• can handle certain finite range dependencies	• assumes independence
• up to 10 x more computation	• no additional computation after sampling from posterior