



Classifiers, denoted as  $f(x; p, D^{*k})$ , are constructed for each candidate number  $p$  of top genes and each PNB dataset  $D^{*k}$ . The evaluated performances of the classifiers  $G(p; D^{*k})$  for  $k = 1, \dots, B$  then simulates the distribution of prediction performances. We consider the lower 10% point as a safe representative of the distribution arising from the observation noise. This is called a Percentile10 classifier candidate from PNB datasets, Percentile10  $\{G(p; D^*)\}$ .

The Percentile method selects the best classifier which has the highest accuracy among percentile10 classifier candidates for different numbers of candidate genes, such as  $p = 2, 4, 10, \dots$ :

$$\operatorname{argmax}_p \text{Percentile10} \{G(p; D^*) | p = 2, 4, 10, \dots\}. \quad (2)$$

### 3 Result

We applied the PNB method to a real dataset genes expressed in neuroblastoma (NBL) [2]. According to multiple observations, we assume the observation noise distribution is Gaussian with variance  $0.3^2$ . In Figure 1, we show the estimated leave-one-out (LOO) cross validation accuracies of WV with various numbers of top genes and their distributions simulated by PNB. Although the classifier with the highest original NBL accuracy was obtained at #gene=29 (where the solid line reached maximum), this point seems to be fortuitous because of small sample numbers and/or observation noise. The PNB method suggests that there is high risk of selecting classifiers with extremely low accuracy when a small  $p$  is selected, but that risk becomes low for a large  $p$ . At #gene=45, where the lower 10% point is maximal, we may expect a good balance between low risk and high expected accuracy.



Figure 1: NBL prognosis at 24months

Solid line shows LOO accuracies obtained from the original NBL dataset. Dashed lines (from bottom) are 10, 50 and 90 percentiles of accuracy distribution simulated by our PNB method.

### References

- [1] Golub, T. R. et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531–537.
- [2] Ohira, M. et al. (2005) Expression profiling using a tumor-specific cDNA microarray predicts the prognosis of intermediate risk neuroblastomas. *Cancer Cell* 7:337–350.