

HIGH-DIMENSIONAL REGRESSION IN CANCER GENOMICS

by Chris Hans and Mike West

hans@stat.ohio-state.edu

mw@isds.duke.edu

Cancer-genomic data present modeling challenges due to their high-dimensional nature. A typical dataset consists of several (tens or hundreds, typically the former) tumor samples which are then used to generate tens of thousands of estimated gene expression levels. Associated with each patient is a particular outcome of interest, e.g. survival time, presence/absence of lymph node invasion, or response to a particular treatment. Two goals are the identification of the complex multivariate patterns of association between key genes and the outcome, and prediction for future patients.

Shotgun Stochastic Search

Because the sample size is often much smaller than the total number of candidate predictor variables, we focus on generating lists of *sparse* regression models — models comprised of only a few genes. By scoring each model by its unnormalized posterior probability, we can build lists with potentially millions of models that exhibit good fit to the data, and then perform model averaging to identify dominant genes and to provide predictive distributions. Emphasis on sparse models is made possible through the use of prior distributions that highly penalize the addition of predictor variables to a model. Indeed, sparsity is emerging as an important component in modeling such high-dimensional datasets.

In order to perform the search necessary to construct such model lists, we introduce the “shotgun stochastic search” (SSS), a parallel-computing based method for exploring large model spaces (see Hans, Dobra and West (2005) for details). SSS is related to MCMC approaches for traversing regression model spaces, however due to the use of a distributed computing environment, SSS is able to more quickly explore local regions of model space, rapidly identifying promising models and providing new directions in which the search may evolve.

Brain Cancer Survival Study

We analyzed gene expression data from a survival study in brain cancer based at the W.M. Keck Center for Neuro-Oncology at Duke University. A detailed description of the data along with an ini-

tial analysis can be found in Rich *et al.* (2005). The study consists of 41 patients diagnosed with glioblastoma, a form of brain cancer associated with relatively short survival times. Although survival times are generally short, significant variation is observed and it is of interest to explore possible biological explanations for this variability by analyzing the gene expression data.

For each patient we have survival time (in days) measured from initial diagnosis along with a tumor specimen. Due to the nature of the disease, all of the patients in the study are deceased and hence there is no censoring information. Gene expression data is available on Affymetrix human U133A microarrays, processed using the current standard RMA method to provide expression estimates for each gene. After an initial screening to remove probes whose expression levels were clearly “in the noise”, a total of 8,408 genes were included in the analysis.

We used SSS to explore the model space of sparse regression models: $\log(\text{survival time}) \sim N(X_\gamma \beta_\gamma, \sigma^2)$, where γ represents a particular (small) subset of genes. The best one million models found by SSS identified several genes that were associated with variability in survival time (see Rich *et al.* (2005) for details). In particular, one gene known to have increased expression levels in several other types of cancer was found to be influential in discriminating survival time, especially in the context of regression models including two other genes that have specific neural functions. The top panel of Figure 1 displays model averaged fitted values with 95% credible intervals.

Versatility of SSS

While this application concerned the normal linear model, SSS can be used to search any discrete space. For example, we performed a separate analysis of the brain cancer data using Weibull regression models with survival time in months as the outcome. A plot of leave-one-out cross-validated predictions of one year survival probability is given in the second panel of Figure 1 — combining information across thousands of Weibull regression models seems to well-discriminate twelve month survival (approximately the general population median). We have also applied SSS to binary regression models (Dressman *et al.*, 2006) and to Gaussian graphical models (Jones *et al.*, 2005), both in contexts of high-dimensional cancer genomic datasets. We anticipate applications in other complicated modeling contexts designed to coherently extract information from these inherently high-dimensional problems.

References

- Dressman, H., Hans, C., Bild, A., Olson, J., Rosen, E., Marcom, P., Liotcheva, V., Jones, E., Vujaskovic, Z., Marks, J., Dewhirst, M., West, M., Nevins, J. and Blackwell, K. (2006). “Gene expression profiles of multiple breast cancer phenotypes and response to neoadjuvant chemotherapy.” *Clinical Cancer Research*, **12**, 819–826.
- Hans, C., Dobra, A. and West, M. (2005). “Shotgun stochastic search for ‘large p ’ regression.” ISDS Discussion Paper 2005-10 (Submitted).
- Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. and West, M. (2005). “Experiments in stochastic computation for high-dimensional graphical models.” *Statistical Science*, **20**, 388–400.
- Rich, J., Hans, C., Jones, B., Iversen, E., McClendon, R., Rasheed, B., Dobra, A., Dressman, H., Bigner, D., Nevins, J. and West, M. (2005). “Gene expression profiling and genetic markers in glioblastoma survival.” *Cancer Research*, **65**, 4051–4058.

Figure 1: The top panel displays model-averaged fitted log survival time vs. observed log survival time (in days) under the normal linear model setup. The bottom panel displays leave-one-out cross-validated model-averaged predicted twelve month survival probabilities for the Weibull regression model; red (blue) points indicate patients with actual survival time less (greater) than one year.

