

BFRM: SOFTWARE FOR BAYESIAN FACTOR REGRESSION MODELS

by Quanli Wang,

Carlos M. Carvalho, Joe Lucas & Mike West
{quanli, carlos, jel2, mike}@stat.duke.edu

Overview

BFRM is a comprehensive implementation of sparse statistical models for high-dimensional data analysis, structure discovery and prediction.

The framework of sparse latent factor modelling coupled with sparse regression and ANOVA for multivariate data is relevant in many exploratory and predictive problems with high-dimensional multivariate observations. Bayesian analysis utilising sparsity-inducing models, and computational methods able to efficiently explore and fit large-scale models, now allow these approaches to be used in increasingly complex and high-dimensional problems.

The statistical methods and computational analysis represented in BFRM are generic and suited to many areas of application. A range of recent applications – and a core set of motivating problems for some of the recent modelling and computational developments – are biological studies using gene expression data. A number of these studies are represented in examples in the papers below. These illustrate exploratory and predictive analyses of gene expression data coupled with outcomes (phenotypes) to be predicted, and related studies in biological pathway analysis.

The main methodological aspects of BFRM are described in Carvalho *et al.* (2007) [1]. Sparse factor modelling developments there build on and develop earlier ideas and methods from West (2003) [4]. BFRM is written in C++ and freely available to interested researchers. The BFRM executables for multiple platforms and operating systems, together with detailed descriptions for installing and running the code and a number of examples, are available at:

<http://xpress.isds.duke.edu:8080/bfrm/>

Examples and Case Studies

The examples in [1], [2] and [3] illustrate the use of BFRM in the following case studies:

[1] Based on the analogy of latent factors representing biological “subpathways” structure, this paper explores connections between factors and multiple biological aspects of cancer genomics. The studies discuss the discovery

use of this approach in expanding the existing knowledge of oncogenic pathways along with the illustration of the predictive ability of aggregate patterns of gene expression profiles in prognostic clinical contexts.

- [2] This paper discusses the use of sparse anova models in gene expression experiments designed to investigate the transcriptional responses to interventions that up-regulate a series of key oncogenes, and includes a number of practical model developments relevant to modern gene expression array technologies.
- [3] This paper describes case studies that use BFRM for the creation of gene expression signatures of cardiovascular disease states, linking to risk factors from designed experiments in mice models. Analysis investigates cross-species extrapolation of risk signatures by projection to human observational data with the latter modelling via sparse latent factor analysis using BFRM.

Example summaries from the cardiovascular genomics studies are in Figure 1. Exploration of these and other graphical and numerical summaries – in designed experiments and observational data sets from cancer and cardiovascular genomics – are developed in the referenced papers.

References

- [1] Carvalho, C., Chang, J., Lucas, J., Wang, Q., Nevins J. and West, M. (2006). “High-dimensional sparse factor modelling: Applications in gene expression genomics.” (Submitted). <http://ftp.stat.duke.edu/WorkingPapers/05-15.html>
- [2] Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins J. and West, M. (2006). “Sparse statistical modelling in gene expression genomics.” In *Bayesian Inference for Gene Expression and Proteomics*, (eds. K.A. Do *et al*), CUP, 155-176. <http://ftp.stat.duke.edu/WorkingPapers/06-01.html>
- [3] Seo, D., Goldschmidt-Clermont P. and West, M. “Of mice and men: Sparse statistical modelling in cardiovascular genomics.” (2007). *Annals of Applied Statistics* **1**, <http://ftp.stat.duke.edu/WorkingPapers/07-05.html>
- [4] West, M. “Bayesian factor regression models in the “large p, small n” paradigm.” (2003). *Bayesian Statistics* **7**, 723-732. <http://ftp.isds.duke.edu/WorkingPapers/02-12.html>

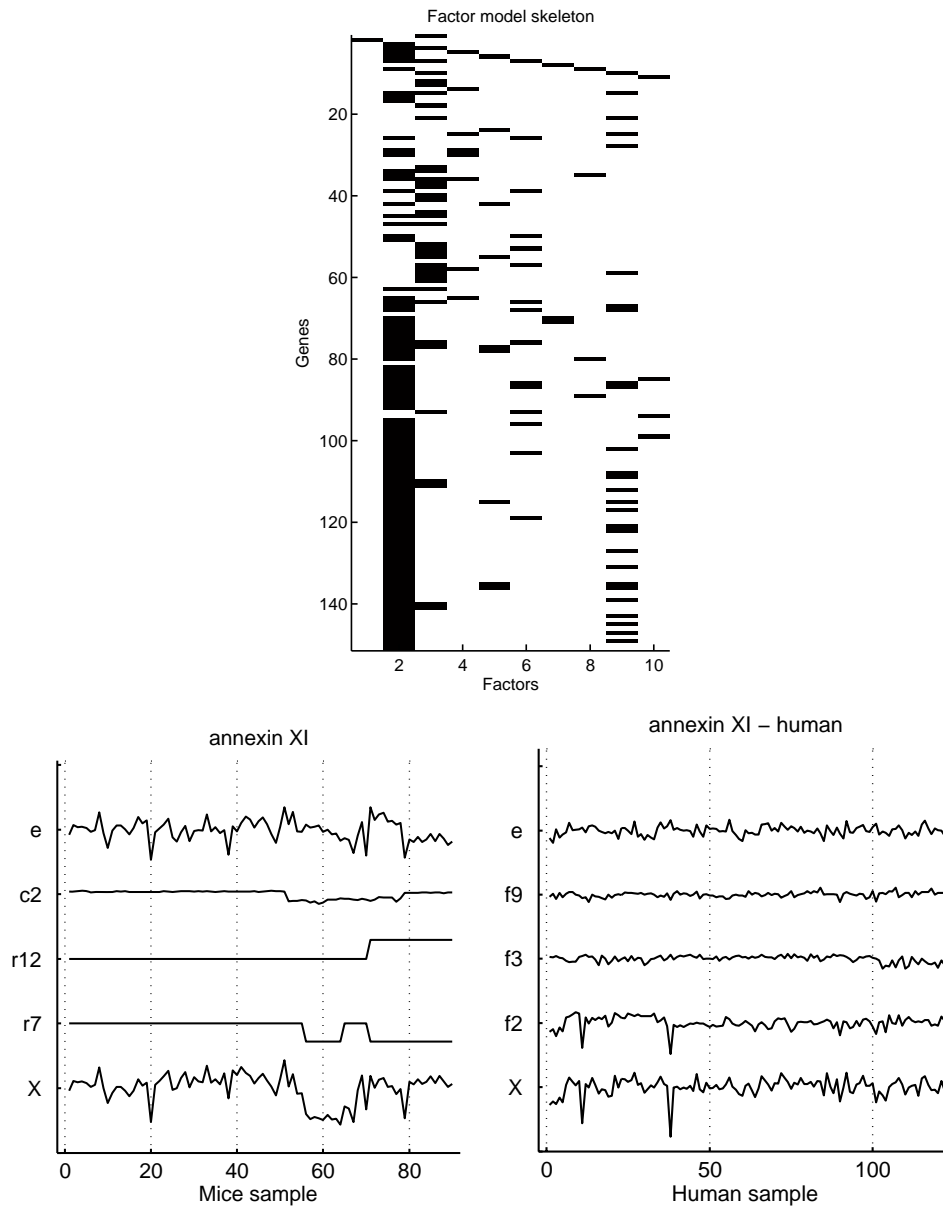


Figure 1: Examples from BFRM analyses of gene expression in designed experiments on mice coupled with human observational data, developed in a cardiovascular genomics project. The top frame is a “skeleton” of a latent factor model – fitted to 150 genes (rows) and involving 10 latent factors (columns) – in the human cardiovascular study, showing the sparsity pattern of gene-factor associations based on thresholding posterior probabilities of such associations (black=“on”, white=“off”). The lower left frame shows a BFRM estimated decomposition of gene expression (X) of mouse gene annexin XI across mouse samples, illustrating estimated contributions to expression of this gene related to two design variables ($r7$, $r12$), a microarray experimental artifact correction regressor ($c2$), and the residual (e). The lower right frame shows a similar decomposition of the same gene but now assayed in the human data, and where the expression fluctuations across human samples relate to estimated latent factors 2,3, and 9.