Variational Bayes for High-dimensional Survival Analysis

CMS 2021

Michael Komodromos

Supervised by: Marina Evagelou, Kolyan Ray and Sarah Filippi

Department of Mathematics
Imperial College London

December 17, 2021

Outline

Summary

Survival Analysis

$$S(t) = 1 - F(t)$$

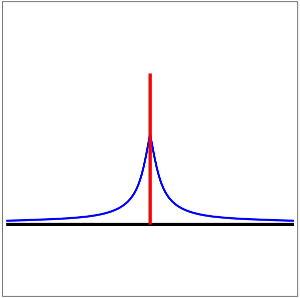
$$S(t) = \exp(-\int h(t)\ dt)$$

$$h(t) = h_0(t)\exp(\beta^\top x)$$

Spike & Slab Priors

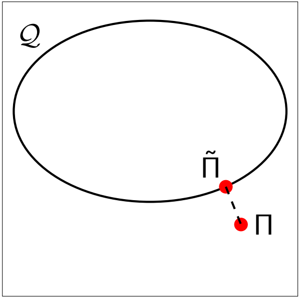
$$\beta_j | z_j \overset{\text{ind}}{\sim} z_j \Psi_j + (1 - z_j)\delta_0$$

$$z_j \overset{\text{iid}}{\sim} \text{Bern}(w)$$



Variational Inference

$$\tilde{\Pi} = \underset{Q \in \mathcal{Q}}{\arg\min}\ \text{KL}\left(Q \| \Pi\right)$$



A bit of biology

High-throughput sequencing, produces large-scale datasets describing the:

- Genome (DNA)
- Transcriptome (RNA)
- Proteome (proteins)
- ...

Motivation

Sequencing gives us incredible opportunities to learn about the biology driving the expression of phenotypes.

And beyond that, clinical phenotypes such as survival times or time to disease.

Deepens our understanding of disease, but also allows us to improve prognosis / biomarker characterization.

Problem

- BUT -

These datasets are massive ($p \gg n$), and therefore computationally and statistically challenging to analyze.

Particularly if we want to do:

- Variable selection
- Effect estimation (+uncertainty quantification)
- Computationally scalable

Survival Analysis

Survival analysis

Let $\boldsymbol{T}$ denote a time to failure event with CDF $F(t), t \in \mathbb{R}^+$.

Survivor function, prob. surviving past time t

$$S(t) = 1 - F(t)$$

Hazard rate, instantaneous rate of failure

$$h(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = -(\log S(t))'$$

[Cla+03; Bra+03]

Survival analysis

Re-arranging gives,

$$S(t) = \exp\left(-\int_0^t h(s)ds\right) \tag{1}$$

We can now express $F(t), S(t), f(t)$ in terms of the hazard function and perform inference

- BUT -

- Survival times are often (right) censored.
- $h(t)$ often requires us to estimate a baseline function

Proportional hazards model

Proportional hazards model

Assume

$$h(t; h_0, \beta, x) = h_0(t) \exp\left(\beta^\top x\right)$$

where $x \in \mathbb{R}^p$ are the predictors, $\beta \in \mathbb{R}^p$ the model coefficients, and $h_0 : \mathbb{R}^+ \to \mathbb{R}$ is the baseline hazard rate (often left unspecified).

Notation: for $i = 1, \ldots, n$ observations

- $t_i \in \mathbb{R}^+$ observed time
- $\delta_i = \mathbb{I}(\text{event has occurred})$
- $x_i \in \mathbb{R}^p$ predictors.

Proportional hazards model

Lets us write down the likelihood as

$$L_p(\mathcal{D};\beta) = \prod_{\{i:\delta_i=1\}} \frac{\exp\left(\beta^\top x_i\right)}{\sum_{r\in R(t_i)} \exp\left(\beta^\top x_r\right)} \tag{2}$$

where $\mathcal{D} = \{(t_i, \delta_i, x_i)\}_{i=1}^n$ and $R(t_i) = \{r : t_r \geq t_i\}$ (aka *risk set*).

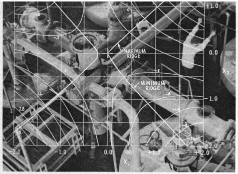
Can be viewed as a profile likelihood - OR - the marginal likelihood assuming a non-informative Gamma process prior over $H_0(t) = \int h_0(t)dt$.

[Cox72; MV00; ICS01]

Spike-and-Slab (SpSL) priors

A brief history

[Hoe62; MB88; GM93]



Application of ridge analysis to regression problems

This new computational procedure is applicable for analyzing regression-type problems in cases involving poorly-conditioned data. Near optimum regression coefficients can be estimated.

Ridge regression

1962

1988

Bayesian Variable Selection in Linear Regression

T. J. MITCHELL and J. J. BEAUCHAMP*

This article is concerned with the selection of subsets of predictor variables in a linear regression model for the prediction of a dependent variable. It is based on a Bayesian approach, intended to be as objective as possible. A probability distribution is first assigned to the dependent variable through the specification of a family of prior distributions for the unknown parameters in the regression model. The method is not fully Bayesian, however, because the ultimate choice of prior distribution from

Spike-and-Slab priors

1993

Variable Selection Via Gibbs Sampling

EDWARD I. GEORGE and ROBERT E. McCULLOCH*

A crucial problem in building a multiple regression model is the selection of predictors to include. The main thrust of this article is to propose and develop a procedure that uses probabilistic considerations for selecting promising subsets. This procedure entails embedding the regression setup in a hierarchical normal mixture model where latent variables are used to identify subset choices. In this framework the promising subsets of predictors can be identified as those with higher posterior probability. The computational burden is then alleviated by using the Gibbs sampler to indirectly sample from this multinomial posterior distribution on the set of possible subset choices. Those subsets with higher probability—the promising ones—can then be identified by their more frequent appearance in the Gibbs sample.

Spike-and-Slab w/ Gibbs

[Tib96; CPS10; RG18]

1996

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]

SUMMARY

We propose a new method for estimation in linear models. The 'lasso' minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less

LASSO

2010

The horseshoe estimator for sparse signals

BY CARLOS M. CARVALHO, NICHOLAS G. POLSON

Booth School of Business, University of Chicago, Chicago, Illinois 60637, U.S.A.

carlos.carvalho@chicagobooth.edu nicholas.polson@chicagobooth.edu

Horseshoe prior

2018

The Spike-and-Slab LASSO

Veronika Ročková[a] and Edward I. George[b]

[a]Department of Econometrics and Statistics at the Booth School of Business of the Unive
University of Pennsylvania, Philadelphia, PA

Spike-and-Slab LASSO

Spike-and-Slab prior

Spike-and-Slab prior

$$\beta_j | z_j \overset{\text{ind}}{\sim} z_j \text{Laplace}(\lambda) + (1 - z_j)\text{Dirac}_0$$

$$z_j | w_j \overset{\text{ind}}{\sim} \text{Bernoulli}(w_j)$$

$$w_j \overset{\text{iid}}{\sim} \text{Beta}(a_0, b_0)$$

Each coefficient β_j has a corresponding latent variable z_j

z_j indicates whether the coefficient takes a value of 0 or not i.e. has an effect on our response and is included in our model

Posterior

$$\Pi(\beta, z, w|\mathcal{D}) \quad \propto \quad L_p(\mathcal{D}|\beta, z, w) \quad \times \quad \Pi(\beta, z, w) \tag{3}$$

The posterior is a rich mathematical object, giving insight into:

- Different possible models
- Coefficients of these models
- A mechanism for variable selection via the posterior inclusion probabilities.

[OS09; OYM17; BCG21]

Practical concerns

- BUT -

Computing the posterior is infeasible even for moderate values of p. Because we have 2^p models to explore.

Common to make computational relaxations, wherein the discrete latent variable z_j is replaced by a continuous random variable taking values in $[0, 1]$, known as *continuous shrinkage priors*.

Often maximum *a posteriori* estimates are returned for β.

Variational Inference

Variational Inference

Variational Inference

Approximate the posterior using a tractable distribution,

$$\tilde{\Pi} = \underset{Q \in \mathcal{Q}}{\operatorname{argmin}} \, \mathrm{KL}\left(Q \parallel \Pi(\cdot|\mathcal{D})\right) \tag{4}$$

where $\mathcal{Q}$ is a tractable family of distributions, known as the *variational family*.

- ✓ Scalable
- ✓ Good point estimates
- ✓ Uncertainty quantification, quality depends on $\mathcal{Q}$

[BKM17; Zha+19]

Variational Family

Variational family

$$\mathcal{Q} = \left\{ Q_{\mu,\sigma,\gamma} = \bigotimes_{j=1}^{p} \left[\gamma_j N(\mu_j, \sigma_j^2) + (1-\gamma_j)\delta_0\right] \right\} \tag{5}$$

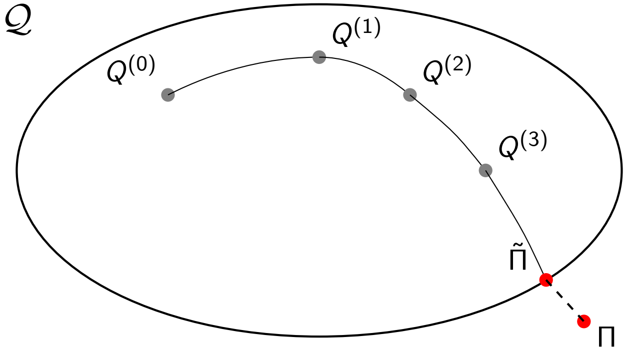
where $\mu_j \in \mathbb{R}, \sigma_j \in \mathbb{R}^+, \gamma_j \in [0,1], j = 1, \ldots, p$. And the notation $\otimes$ is the product measure.

In effect

$$\beta_j \overset{\text{ind}}{\sim} \gamma_j N(\mu_j, \sigma_j^2) + (1-\gamma_j)\delta_0$$

Finding the variational posterior

Convenient to use co-ordinate ascent variational inference, parameters $\mu_j, \sigma_j, \gamma_j$ are updated sequentially keeping the rest fixed.



VI: Practical concerns

Solving the objective is often non-convex, therefore can be sensitive to the starting value.

In practice we found good starting values often yield better results.

Simulations & Application

Simulations

- $n = 200$
- $p = 1000$
- Censoring proportion of $c = 0.25$ or 0.4
- True β_0 with 10 non-zero values sampled uniformly from $[-2, -0.5] \cup [0.5, 2.0]$
- Predictors from one of:
  - *Setting 1*: $x_{ij} \overset{\text{iid}}{\sim} N(0, 1)$
  - *Setting 2*: $x_i \overset{\text{iid}}{\sim} N(0, \Sigma)$, predictors are moderately correlated within groups and not them.
  - *Setting 3*: x_i sampled without replacement from a real dataset.

Comparison to MCMC: results

c	Method	ℓ_2^2-error	ℓ_1-error	TPR	FDR	AUC	runtime
0.25	VB	0.196 (0.177)	1.098 (0.477)	0.993 (0.026)	0.000 (0.000)	0.999 (0.005)	24.7s (6.3s)
	MCMC	0.224 (0.200)	1.141 (0.506)	0.990 (0.033)	0.000 (0.000)	0.999 (0.005)	4h 4m (2h 22m)
0.4	VB	0.277 (0.255)	1.272 (0.588)	0.980 (0.051)	0.001 (0.009)	0.996 (0.015)	20.6s (4.7s)
	MCMC	0.361 (0.361)	1.425 (0.704)	0.975 (0.056)	0.001 (0.009)	0.998 (0.009)	4h 54m (2h 12m)
0.25	VB	0.528 (0.702)	1.633 (1.137)	0.948 (0.085)	0.031 (0.074)	0.981 (0.033)	22.6s (5.7s)
	MCMC	0.428 (0.493)	1.487 (0.869)	0.951 (0.087)	0.004 (0.022)	0.995 (0.018)	4h 4m (2h 14m)
0.4	VB	0.722 (0.833)	1.936 (1.240)	0.921 (0.102)	0.031 (0.064)	0.971 (0.040)	20.2s (5.2s)
	MCMC	0.899 (1.571)	2.089 (1.649)	0.900 (0.160)	0.008 (0.031)	0.991 (0.024)	4h 36m (3h 39m)
0.25	VB	5.752 (3.254)	5.769 (2.192)	0.601 (0.174)	0.053 (0.109)	0.852 (0.081)	14.7s (6.6s)
	MCMC	5.750 (2.847)	5.746 (1.899)	0.577 (0.184)	0.016 (0.069)	0.881 (0.069)	4h 45m (2h 45m)
0.4	VB	7.390 (4.001)	7.007 (2.573)	0.497 (0.210)	0.060 (0.130)	0.805 (0.089)	7.7s (2.8s)
	MCMC	7.400 (3.435)	6.870 (2.134)	0.482 (0.199)	0.024 (0.087)	0.849 (0.079)	2h 28m (55m 9s)

Comparison to MCMC: uncertainty quantification

c	Method	Cvrg. $\beta_0 \neq 0$	Set size $\beta_0 \neq 0$	Cvrg. $\beta_0 = 0$	Set size $\beta_0 = 0$
0.25	VB	0.770 (0.202)	0.320 (0.013)	1.000 (0.000)	0.000 (0.000)
	MCMC	0.928 (0.138)	0.506 (0.039)	1.000 (0.000)	0.000 (0.000)
0.4	VB	0.774 (0.208)	0.355 (0.021)	1.000 (0.000)	0.000 (0.000)
	MCMC	0.914 (0.127)	0.570 (0.054)	1.000 (0.000)	0.000 (0.000)
0.25	VB	0.703 (0.227)	0.306 (0.028)	1.000 (0.001)	0.000 (0.000)
	MCMC	0.904 (0.161)	0.522 (0.053)	1.000 (0.000)	0.000 (0.000)
0.4	VB	0.683 (0.262)	0.333 (0.039)	1.000 (0.001)	0.000 (0.000)
	MCMC	0.845 (0.218)	0.567 (0.101)	1.000 (0.000)	0.000 (0.000)
0.25	VB	0.427 (0.205)	0.316 (0.099)	1.000 (0.001)	0.000 (0.001)
	MCMC	0.529 (0.210)	0.431 (0.145)	1.000 (0.000)	0.000 (0.000)
0.4	VB	0.342 (0.208)	0.276 (0.123)	1.000 (0.001)	0.000 (0.001)
	MCMC	0.436 (0.220)	0.400 (0.176)	1.000 (0.000)	0.000 (0.000)

Comparison to other methods

Compare against other Bayesian PHM variable selection methods

- **BhGLM** spike-and-slab LASSO method
- **BVSNLP** inverse moment prior with Dirac spike

both return MAP estimates for β and inclusion probabilities.

Changes

- $n = 1000$ (bar setting 3 where it's 500)
- $p = 10,000$
- 60 non-zero values in β_0

Comparison to other methods

c	Method	ℓ_2^2-error	ℓ_1-error	TPR	FDR	AUC
	SVB	**0.216 (0.172)**	**2.834 (1.135)**	**1.000 (0.000)**	**0.000 (0.000)**	**1.000 (0.000)**
0.25	BhGLM	12.183 (2.361)	36.836 (2.511)	**1.000 (0.000)**	**0.000 (0.000)**	**1.000 (0.000)**
	BVSNLP	0.977 (6.533)	3.382 (3.428)	**1.000 (0.005)**	**0.000 (0.000)**	**1.000 (0.002)**
	SVB	**0.327 (0.250)**	**3.510 (1.449)**	**1.000 (0.000)**	**0.000 (0.000)**	**1.000 (0.000)**
0.4	BhGLM	6.239 (1.768)	26.806 (2.774)	**1.000 (0.000)**	**0.000 (0.000)**	**1.000 (0.000)**
	BVSNLP	29.117 (48.539)	22.898 (32.593)	0.760 (0.406)	0.003 (0.015)	0.896 (0.177)
	SVB	**0.221 (0.156)**	**2.857 (1.018)**	**1.000 (0.000)**	**0.000 (0.002)**	**1.000 (0.000)**
0.25	BhGLM	3.089 (0.987)	19.276 (2.312)	**1.000 (0.000)**	**0.000 (0.000)**	**1.000 (0.000)**
	BVSNLP	0.238 (0.119)	2.953 (0.718)	**1.000 (0.000)**	**0.000 (0.000)**	**1.000 (0.000)**
	SVB	**0.340 (0.236)**	**3.586 (1.348)**	**1.000 (0.000)**	**0.000 (0.002)**	**1.000 (0.000)**
0.4	BhGLM	1.568 (0.636)	13.654 (2.107)	**1.000 (0.002)**	**0.000 (0.000)**	**1.000 (0.000)**
	BVSNLP	7.100 (25.053)	8.370 (18.197)	0.947 (0.200)	0.020 (0.079)	0.977 (0.088)
	SVB	**88.538 (13.392)**	**71.139 (6.296)**	0.202 (0.085)	**0.375 (0.163)**	0.608 (0.044)
0.25	BhGLM	97.553 (24.609)	83.879 (14.873)	**0.224 (0.143)**	0.602 (0.247)	**0.618 (0.072)**
	BVSNLP	96.940 (13.788)	74.729 (5.890)	0.173 (0.076)	0.499 (0.127)	0.604 (0.038)
	SVB	**93.753 (12.268)**	**72.888 (5.421)**	**0.149 (0.071)**	**0.388 (0.174)**	**0.581 (0.036)**
0.4	BhGLM	105.312 (19.697)	86.199 (11.064)	**0.149 (0.102)**	0.674 (0.216)	0.579 (0.053)
	BVSNLP	100.738 (12.474)	75.886 (4.835)	0.123 (0.055)	0.526 (0.139)	0.579 (0.031)

Ovarian Cancer Transcriptomics Dataset

Dataset describing the genes expressed in tumors of patients with ovarian cancer

- $n = 580$ with 39.5% censored
- $p = 12,042$

Aim: identify which genes are associated with overall survival

We fit models fixing $a_0 = p/100$ and $b_0 = p$ and considered different values of λ

OvC: Results

We compute the selection proportion of each gene, i.e. the number of times across the different models a gene had a posterior inclusion probability greater than 0.5.

PI3	VSIG4	PPP3CA	IL7R	SDF2L1	D4S234E	DAP	CCR7
0.786	0.307	0.257	0.243	0.207	0.2	0.193	0.186
ACSL3	PLA2G2D	ADORA3	FLNA	SLAMF7	UBD	CD14	HABP2
0.157	0.157	0.121	0.121	0.107	0.107	0.086	0.086
LPXN	LCE2B	TBP	GALNT10	NOTCH4	RNF128	C5orf28	PPM2C
0.086	0.079	0.079	0.071	0.071	0.071	0.064	0.064
FJX1	TSPAN13	HSPB7	TREML2				
0.057	0.057	0.05	0.05				

Genes with biological interpretations discovered in the biomedical literature

Resources

Paper

Currently in submission

Variational Bayes for survival

https://github.com/mkomod/survival.svb

MCMC sampler

https://github.com/mkomod/survival.ss

Slides

https://github.com/mkomod/presentations

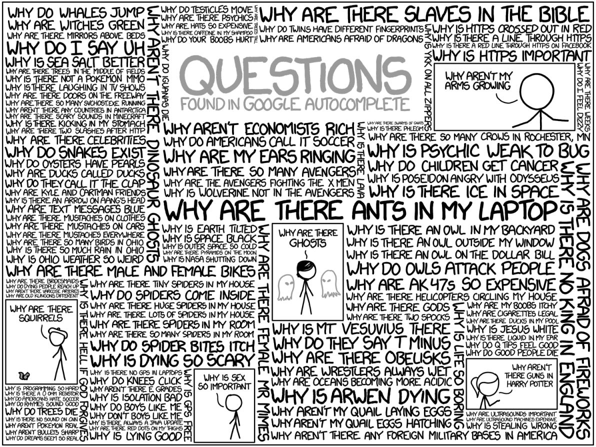


Figure: `https://xkcd.com/1256/`

Reference I

[BCG21] Sayantan Banerjee, Ismaël Castillo, and Subhashis Ghosal. *Bayesian inference in high-dimensional models*. 2021.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. Chap. 10.

[BKM17] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877.

[Bra+03] M. J. Bradburn et al. "Survival Analysis Part II: Multivariate data analysis- An introduction to concepts and methods". In: *British Journal of Cancer* 89.3 (2003), pp. 431–436.

Reference II

[Cla+03] T. G. Clark et al. "Survival Analysis Part I: Basic concepts and first analyses". In: *British Journal of Cancer* 89.2 (2003), pp. 232–238.

[Cox72] David R Cox. "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society, Series B* 34.2 (Feb. 1972), pp. 187–220.

[CPS10] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. "The horseshoe estimator for sparse signals". In: *Biometrika* 97.2 (2010), pp. 465–480.

Reference III

[CS12] Peter Carbonetto and Matthew Stephens. "Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies". In: *Bayesian Analysis* 7.1 (2012), pp. 73–108.

[GM93] Edward I. George and Robert E. McCulloch. "Variable Selection via Gibbs Sampling". In: *Journal of the American Statistical Association* 88.423 (1993), pp. 881–889.

[Hoe62] A.E. Hoerl. "Application of ridge analysis to regression problems". In: *Chemical Engineering Progress* 58.3 (1962), pp. 54–59.

[ICS01] Joseph G. Ibrahim, Ming-Hu Chen, and Debajyoti Sinha. *Bayesian Survival Analysis*. Springer, 2001, pp. 53–55.

Reference IV

[MB88] T. J. Mitchell and J. J. Beauchamp. "Bayesian variable selection in linear regression". In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1023–1032.

[MV00] S. A. Murphy and A. W. Van Der Vaart. "On profile likelihood". In: *J. Am. Stat. Assoc.* 95.450 (2000), pp. 449–465.

[OS09] R. B. O'Hara and M. J. Sillanpää. "A review of bayesian variable selection methods: What, how and which". In: *Bayesian Anal.* 4.1 (2009), pp. 85–118.

[OYM17] John T. Ormerod, Chong You, and Samuel Müller. "A variational bayes approach to variable selection". In: *Electron. J. Stat.* 11.2 (2017), pp. 3549–3594.

Reference V

[RG18] Veronika Ročková and Edward I. George. "The Spike-and-Slab LASSO". In: *Journal of the American Statistical Association* 113.521 (2018), pp. 431–444.

[Tib96] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society, Series B* 58.1 (1996), pp. 267–288.

[Zha+19] Cheng Zhang et al. "Advances in Variational Inference". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 41.8 (2019), pp. 2008–2026.