Default Priors and Robust Estimation for Generalized Linear Models

(A.K.A., A few things I learned from Luis Pericchi)

Abel Rodríguez

O'Bayes Meeting
April, 2022

UNIVERSITY of WASHINGTON

Department of STATISTICS

Outline

Once upon a time ...

Training samples

Prior matching

Heavy tail priors

Concluding remarks

"Training" your improper prior

- The idea of splitting your sample into a training set and a test set is at the core of statistics and machine learning.
- Use the training set, y^*, to turn your improper prior into a proper one, and the rest, y^{**}, to run the actual test.

$$\pi_k^I(\boldsymbol{\theta}_k \mid M = k, \mathrm{y}^*) = \frac{p_k(\mathrm{y}^* \mid \boldsymbol{\theta}_k, M = k)\pi_k^N(\boldsymbol{\theta}_k \mid M = k)}{m_k^N(\mathrm{y}^*)}$$

$$m_k^I(\mathrm{y}^{**} \mid M = k, \mathrm{y}^*) = \int p_k(\mathrm{y}^{**} \mid \mathrm{y}^*, \boldsymbol{\theta}_k, M = k)\pi_k^I(\boldsymbol{\theta}_k \mid M = k, \mathrm{y}^*)\mathrm{d}\theta_k$$

$$B_{k,k'}^I(\mathrm{y}) = \frac{m_k^I(\mathrm{y}^{**} \mid M = k, \mathrm{y}^*)}{m_{k'}^I(\mathrm{y}^{**} \mid M = k', \mathrm{y}^*)}$$

- If you average over training samples $\Rightarrow$ *Intrinsic Bayes Factor* (Berger & Pericchi, 1996).
- Size of the training sample? $\Rightarrow$ Minimal!

EP priors

- Introduced by Pérez & Berger (2002).
- EP priors follow a similar rationale, but use an *imaginary* training sample, which is averaged out!

$$\pi_k^E(\boldsymbol{\theta}_k \mid M_k) = \int \frac{p_k(\mathrm{y}^* \mid \boldsymbol{\theta}_k, M = k)\pi_k^N(\boldsymbol{\theta}_k \mid M = k)}{m_k^N(\mathrm{y}^*)} m^*(\mathrm{y}^*)\,\mathrm{d}\mathrm{y}^*$$

- Choosing the training sample is replaced with choosing m^*.
- For nested models, a common option is $m^*(\mathrm{y}^*) = m_0^N(\mathrm{y}^*)$ ("simplest" model), which makes the EP prior asymptotically equivalent to the prior implied by the AIBF!

PEP priors

- Introduced by Fouskakis et al. 2015.
- Similar to the EP prior, but it "scales" training likelihood to be (approximately) unit information!

$$\pi_k^{PEP}(\boldsymbol{\theta}_k \mid M_k) \propto$$
$$\int \frac{\{p_k(\mathrm{y}^* \mid \boldsymbol{\theta}_k, M = k)\}^{1/\delta}\, \pi_k^N(\boldsymbol{\theta}_k \mid M = k)}{\int \{p_k(\mathrm{y}^* \mid \boldsymbol{\theta}_k, M = k)\}^{1/\delta}\, \pi_k^N(\boldsymbol{\theta}_k \mid M = k) \mathrm{d}\boldsymbol{\theta}_k} m^*(\mathrm{y}^* \mid \delta)\, p(\delta) \mathrm{d}\mathrm{y}^* \mathrm{d}\delta$$

  where $p(\delta)$ has mean n^*.
- Computationally tractable when $p_k(\mathrm{y}^* \mid \boldsymbol{\theta}_k, M = k)$ is Gaussian.

An example: Linear Models

- When $\mathrm{y} \mid \boldsymbol{\theta}_k, \sigma^2, \mathrm{X}_k \sim \mathrm{N}\left(\mathrm{y} \mid \mathrm{X}_k\boldsymbol{\theta}_k, \sigma^2\mathrm{I}\right)$ and $\pi_k^N(\boldsymbol{\theta}_k) \propto 1$ and

$$\pi_k^{PEP}(\boldsymbol{\theta}_k \mid M_k) =$$
$$\int \mathrm{N}\left(\boldsymbol{\theta}_k \mid \left\{\mathrm{X}_k^{*T}\mathrm{X}_k^*\right\}^{-1}\mathrm{X}_k^{*T}\mathrm{y}^*, \delta\sigma^2\left\{\mathrm{X}_k^{*T}\mathrm{X}_k^*\right\}^{-1}\right) m^*(\mathrm{y}^* \mid \delta)\, p(\delta)\mathrm{d}\delta\mathrm{d}\mathrm{y}^*$$

- For $n^* = n$ and $\mathrm{X}_k^* = \mathrm{X}_k$, compare that with the corresponding g-prior:

$$\pi_k^g(\boldsymbol{\theta}_k \mid M_k) = \int \mathrm{N}\left(\boldsymbol{\theta}_k \mid 0, \delta\sigma^2\left\{\mathrm{X}_k^{*T}\mathrm{X}_k^*\right\}^{-1}\right)\tilde{p}(\delta)\mathrm{d}\delta$$

(Recall that $\tilde{p}$ is centered around n.)

Generalizing PEPs to GLMs

- In the case of Gaussian linear models, the PEP is relatively easy to derive because rescaling by $1/\delta$ leaves the likelihood in the normal family.
- The same is not true for other members of the exponential family (e.g., logistic or loglinear regression).
- Fouskakis et al. (2018) propose a generalization, but it has various theoretical, computational and empirical drawbacks.
- We propose a different generalization: the Laplace PEPs!
    - Porwal, A., & Rodriguez, A. (2021). Laplace Power-expected-posterior priors for generalized linear models with applications to logistic regression. arXiv preprint arXiv:2112.02524.

Laplace Power-expected-posterior Prior (LPEP)

- To construct the Laplace PEP, replace $p_k(\mathrm{y}^* \mid \boldsymbol{\theta}_k, M = k)$ with its Laplace approximation **before** raising to the $1/\delta$ power!

$$\pi_k^{PEP}(\boldsymbol{\theta}_k \mid M_k) \propto \int \frac{\mathsf{N}\left(\boldsymbol{\theta}_k \mid \hat{\boldsymbol{\theta}}_k\left(\mathrm{y}^*\right), \delta H_k^{-\mathbf{1}}\left(\hat{\boldsymbol{\theta}}_k\left(\mathrm{y}^*\right)\right)\right) \pi_k^N(\boldsymbol{\theta}_k \mid M = k)}{\int \mathsf{N}\left(\boldsymbol{\theta}_k \mid \hat{\boldsymbol{\theta}}_k\left(\mathrm{y}^*\right), \delta H_k^{-\mathbf{1}}\left(\hat{\boldsymbol{\theta}}_k\left(\mathrm{y}^*\right)\right)\right) \pi_k^N(\boldsymbol{\theta}_k \mid M = k)\mathrm{d}\boldsymbol{\theta}_k} m^*\left(\mathrm{y}^*\right) p(\delta)\mathrm{d}\mathrm{y}^*\mathrm{d}\delta$$

- Laplace approximation should be particularly accurate when $n^* = n$ but, conceptually, the procedure is reasonable for other choices of n^*.
- An implicit constraint on y^* is that the Laplace approximation needs to be well defined (e.g., the MLE needs to exist for every model under consideration).

Example: Logistic regression

- Likelihood is

$$p_k(\mathbf{y} \mid \boldsymbol{\theta}_k, M = k) = \prod_{i=1}^{n} \frac{\exp\left\{ y_i \mathbf{x}_{i,k}^T \boldsymbol{\theta}_k \right\}}{1 + \exp\left\{ \mathbf{x}_{i,k}^T \boldsymbol{\theta}_k \right\}}$$

- Pick $m^*(\mathbf{y}^*) = m_0^N(\mathbf{y}^*) 1(\mathbf{y}^* \in \Omega_k(\mathbf{X}))$ where

$$\Omega_k(\mathbf{X}) = \{\mathbf{y} : \hat{\boldsymbol{\theta}}_k(\mathbf{y}, \mathbf{X}) \text{ is finite for all } k\}$$

 and

$$m_0^N(\mathbf{y}^*) = \frac{\Gamma(y_\cdot^* + 1/2)\Gamma(n - y_\cdot^* + 1/2)}{\Gamma(n+1)\left(\Gamma(1/2)\right)^2}$$

- Various possible choices for $p(\delta)$ (fixed, hyper-g/n, robust prior).

Example: Logistic regression

- For logistic regression (and many other GLMs!) it is enough to show that $\hat{\theta}_k(\mathrm{y}, \mathrm{X})$ is finite when k corresponds to the full model!
  - We provide easy-to-verify sufficient conditions in the paper.
- Checking this condition for logistic regression is relatively straightforward using the algorithm of Kosmidis and Schumacher (2020).
- In this case, the prior is proper for every model k.

Properties of the LPEP for GLMs

- For linear models, this is just your standard PEP!
- Under standard regularity conditions Bayes factors / posterior model probabilities are consistent.
  - True even if p grows with n at a reasonably slow rate.
- Well-defined intrinsic prior.
- Unlike Li & Clyde, (2018), it can be used even if the original data is separable, or in hierarchical settings.
  - Good theoretical properties.
- The fact that they correspond to mixtures of normals facilitates computation using MCMC.
  - For a number of GLMs, no need for reversible-jump schemes like Fouskakis et al. (2018).
  - A second Laplace approximation can be used to speed up computation as in Li & Clyde, (2018).

Simulation study: Design

- $n = 500$; $p = 100$; 100 bootstrapped datasets
- Columns of X drawn from standard normal distribution with pairwise correlation $cor(X_i, X_j) = r^{|i-j|}$ for $1 \leq i < j \leq p$
- Scenarios: $r = 0$ (independent design) and $r = 0.75$ (correlated design)
- $p_{\mathcal{M}_T}$ denote the number of variables in the true model
- $\boldsymbol{b} = (2, -1, -1, 0.5, -0.5)^T$ and $\beta_{\mathcal{M}_T,21:100} = 0$

p_{M_T}	$\beta_{M_T,0}$	$\beta_{M_T,1:5}$	$\beta_{M_T,6:10}$	$\beta_{M_T,11:15}$	$\beta_{M_T,16:20}$
0	-0.5	0	0	0	0
5	-0.5	$\boldsymbol{b}$	0	0	0
10	-0.5	$\boldsymbol{b}$	0	$\boldsymbol{b}$	0
20	-0.5	$\boldsymbol{b}$	$0.5\boldsymbol{b}$	$\boldsymbol{b}$	$0.5\boldsymbol{b}$

- Comparison with: Mixture of g-priors (Li & Clyde, 2018), LASSO, SCAD and MCP.

Simulation study: Results - MAP model properties

p		100							
$p(M)$		Beta-Binomial(1,1)							
p_{M_T}		0		5		10		20	
r		0	0.75	0	0.75	0	0.75	0	0.75
$\delta = n$	LPEP	99	**100***	45	4	**18***	0	0	0
	LCE	**100***	**100***	45	**5**	8	0	0	0
	LCL	**100***	**100***	**46**	4	11	0	0	0
$\delta \sim$ robust	LPEP	99	**100***	**53***	**6***	**15**	0	0	0
	LCE	99	**100***	45	**6***	0	0	0	0
	LCL	**100***	**100***	46	**6***	2	0	0	0
$\delta \sim$ hyper g/n	LPEP	**98**	**100***	**50**	**5**	**17**	0	0	0
	LCE	97	99	25	4	0	0	0	0
	LCL	65	78	3	0	0	0	0	0
	LASSO	59	65	0	0	0	0	0	0
	SCAD	57	59	0	0	0	0	0	0
	MCP	**73**	**66**	**8**	0	**3**	0	0	0

Table: Number of times (**over 100 replications**) that the **MAP** model coincides with the true model in the logistic regression ; **BOLD** represent group maximum; * represent overall maximum.

Simulation study: Results - F1 score

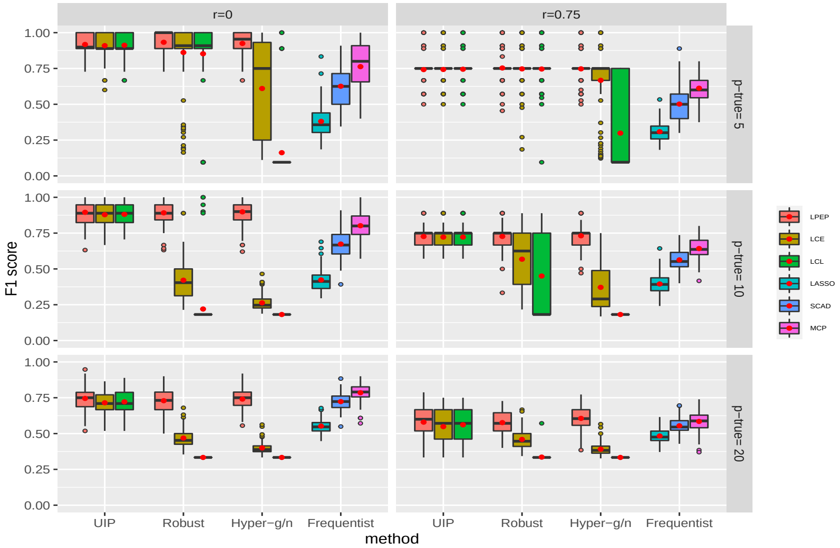


Figure: F1 score across 100 simulated datasets; Red dots represent the average

Simulation study: Results - Mean squared error

p		100							
$p(M)$		Beta-Binomial(1,1)							
p_{M_T}		0		5		10		20	
r		0	0.75	0	0.75	0	0.75	0	0.75
$\delta = n$	LPEP	0.11	**0.10***	2.91	**7.67**	7.09	**17.67**	**14.70**	**33.90**
	LCE	0.11	**0.10***	3.06	7.78	7.64	18.44	16.11	36.47
	LCL	**0.10***	**0.10***	**2.87**	7.68	**6.78**	18.17	16.22	36.43
$\delta \sim$ robust	LPEP	0.12	**0.10***	**2.62***	**6.87***	**6.04***	**14.07***	**13.38**	**24.03***
	LCE	0.12	0.11	4.83	7.80	47.30	23.30	96.14	52.80
	LCL	**0.10***	**0.10***	8.86	8.44	214.63	60.56	275.93	115.58
$\delta \sim$ hyper g/n	LPEP	**0.16**	0.14	**2.70**	**6.89**	**6.12**	**14.76**	**13.03***	**24.86**
	LCE	0.23	**0.13**	6.71	8.90	38.54	26.25	51.48	44.29
	LCL	0.29	0.31	34.28	22.93	104.10	72.95	130.80	94.98
	LASSO	0.25	0.20	7.08	11.91	17.15	25.04	29.44	36.69
	SCAD	**0.21**	**0.16**	3.07	9.02	6.62	**18.80**	14.88	**33.00**
	MCP	0.22	0.18	**2.82**	**8.92**	**6.35**	19.38	**15.13**	33.52

Table: 1000 times the AMSE for estimated coefficients over 100 replications; **BOLD** represent group minimum; * represent overall minimum.

Gusto-I study: survival to treatments for occluded coronary arteries

- Model the binary endpoint of 30-day survival for a subgroup of $n = 2188$ patients using 17 clinical covariates

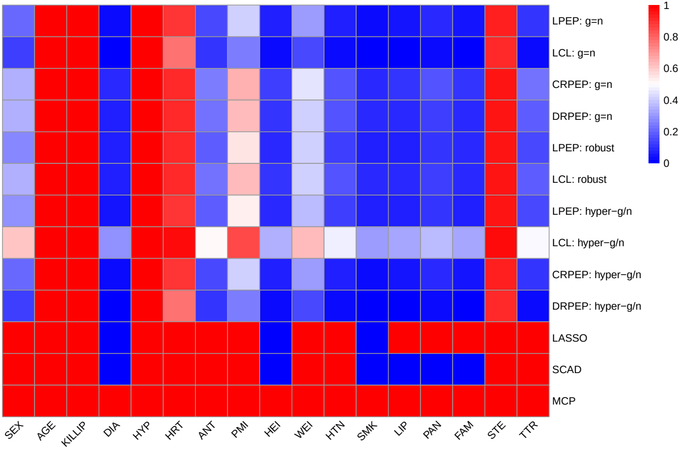


Figure: Marginal posterior inclusion probabilities (PIPs) for `GUSTO-I` dataset (Bayesian procedures) and variables included in the model (penalized likelihood methods).

GUSTO-I study: Out-of-sample Predictive Performance

- We performed a 10-fold cross-validation study
- AUC and Calibration slope (CS) allow us to evaluate the methods in terms of discrimination and calibration; $\uparrow$ score is better
- LS & BRIER measure the predictive accuracy of methods; $\downarrow$ score is better

		AUC	CS	LS	BRIER
$\delta = n$	LPEP	**0.8324***	0.9971	**0.1824**	**0.0496**
	LCL	0.8300	0.9931	0.1831	0.0497
	CRPEP	0.7789	1.0578	0.1965	0.0521
	DRPEP	0.7790	1.0569	0.1963	0.0521
$\delta \sim$ robust	LPEP	**0.8322**	**1.0129**	**0.1822**	**0.0495**
	LCL	0.8316	0.9804	**0.1822**	**0.0495**
$\delta \sim$ hyper g/n	LPEP	**0.8319**	**1.0074***	0.1823	0.0495
	LCL	0.8311	1.0109	**0.1818**	**0.0493**
	CRPEP	0.7956	1.1677	0.1951	0.0522
	DRPEP	0.7800	1.0571	0.1961	0.0520
	LASSO	**0.8305**	**1.0369**	**0.1816***	**0.0492***
	SCAD	0.8243	0.9135	0.1838	0.0496
	MCP	0.8250	0.9196	0.1838	0.0496

Table: Average prediction accuracy measures in a 10-fold cross validation study for `GUSTO-I` dataset

Factor models

- Consider multivariate responses $\mathrm{y}_i = (y_{i,1}, \dots, y_{i,J})^T$ where $y_{i,j} \in \mathbb{R}$ and $i = 1, \dots, I$. A factor model takes the form

$$y_{i,j} = \mu_j + \boldsymbol{\alpha}_j^T \boldsymbol{\beta}_i + \epsilon_{i,j} \qquad\qquad \epsilon_{i,j} \sim \mathrm{N}(0, \sigma_j^2)$$

$\boldsymbol{\alpha}_j^T = (\alpha_{j,1}, \dots, \alpha_{j,d})$, $\boldsymbol{\beta}_i^T = (\beta_{i,1}, \dots, \beta_{i,d})$, and $d \ll J$.

- Used for dimensionality reduction, covariance estimation, prediction.
- The same bilinear structure can be built into Generalized Linear Models. For example, for binary data $y_{i,j} \in \{0, 1\}$,

$$y_{i,j} \sim \mathrm{Ber}\left(\theta_{i,j}\right) \qquad\qquad \theta_{i,j} = G_j\left(\mu_j + \boldsymbol{\alpha}_j^T \boldsymbol{\beta}_i\right)$$

where G is a link function (probit, logit, etc).

- Can be naturally extended to network/relational data.

Factor models: challenges

- Common practical challenges related to model selection:
  - Selecting the dimension d of the latent space.
  - Selecting between a parametric and a non-parametric specification for the distribution of the latent traits.
- The parameters of the model are not identifiable without incorporating some constraints.
  - This can make interpretation and prior elicitation hard.
- Priors need to be chosen very carefully if comparisons are going to be meaningful.

Factor models: selecting d

- Consider a slight generalization of the factor model where

$$y_{i,j} = \mu_j + \boldsymbol{\alpha}_j^T \boldsymbol{\Lambda} \boldsymbol{\beta}_i + \epsilon_{i,j}$$

  where $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_d\}$ and $\lambda_k \in \{0, 1\}$

- The introduction of the λ_ks would in principle enable inference of the dimension of the latent space.

- Note that

$$\text{Var}\left(y_{i,j} \mid \mu_j, \boldsymbol{\alpha}_j, \boldsymbol{\Lambda}\right) = \text{Var}\left(\epsilon_{i,j}\right) + \sum_{k=1}^{d} \lambda_k \alpha_{j,k} \text{Var}\left(\beta_{i,k}\right)$$

- If i.i.d. priors are used for the $\beta_{i,k}$s (which is common), then

$$\lim_{d \to \infty} \text{Var}\left(y_{i,j} \mid \mu_j, \boldsymbol{\alpha}_j, \boldsymbol{\Lambda}\right) = \infty$$

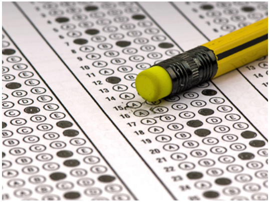
Factor models: selecting d

- There are a couple of possible solutions:
  - Allow the variance of $\beta_{i,k}$ to decrease with k fast enough, for example $\mathrm{Var}\left(\beta_{i,k}\right) = \mathcal{O}(k^{-2})$.
  - Allow $Pr(\lambda_k = 1)$ to decrease fast enough with k.
- This setting extends to factor models embedded in GLMs.
- We have used these approaches in a few papers:
  - Guha, S. & Rodriguez, A. (2021). Bayesian regression with undirected network predictors with an application to brain connectome data. Journal of the American Statistical Association, 116(534), 581-593.
  - Sosa, J. & Rodríguez, A. (2021). A latent space model for cognitive social structures data. Social Networks, 65, 85-97.
  - Guhaniyogi, R. & Rodriguez, A. (2020). Joint modeling of longitudinal relational data and exogenous variables. Bayesian Analysis, 15(2), 477-503.

Factor models: selecting d

Underlying principle: when eliciting priors on non-identifiable parameters for various models, the implied priors on key identifiable quantities should be similar across models.

Factor models: parametric vs. non-parametric specifications



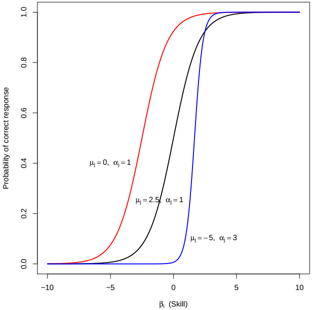
- Consider the 1D factor model:

$$y_{i,j} \sim \text{Ber}\left(G(\mu_j + \alpha_j \beta_i)\right)$$

- Motivating application: *item response models*

i	= Test subject
j	= Question
μ_j	= Difficulty
$\boldsymbol{\alpha}_j$	= Discrimination
$\boldsymbol{\beta}_i$	= Skill

- Rasch model is a special case.



Factor models: parametric vs. non-parametric specifications

- Two possible specifications for the random effect:
  - Standard parametric model: $\beta_i \sim \mathrm{N}(0, 1)$
  - Non-parametric specification (Dirichlet process mixture of normals):

$$\beta_i \mid G \sim \int \mathrm{N}(\cdot \mid \eta, \tau^2) G(\mathrm{d}\eta, \mathrm{d}\tau^2), \qquad G \sim \mathrm{DP}(M, G_0)$$

- How do you fairly compare these two models?
  - Paganin, S., Paciorek, C. J., Wehrhahn, C., Rodriguez, A., Rabe-Hesketh, S., & de Valpine, P. (2022+). Computational methods for Bayesian semiparametric Item Response Theory models. arXiv preprint arXiv:2101.11583.
  - Try to match the prior distribution of $\theta_i = G(\mu_j + \alpha_j \beta_i)$ across both models!

Binary factor models in general topological spaces

- The models we discussed previously project the data on low-dimensional Euclidean spaces.
- In some applications (e.g., in political sciences) other geometries might be more appropriate!

Spatial voting models

Rational choice theory derivation:

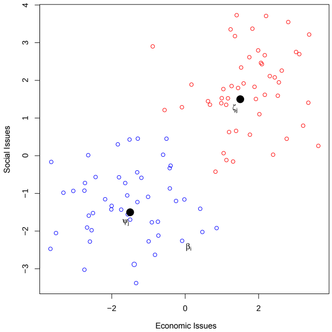
$\boldsymbol{\psi}_j$ = "Yeah" position $\in \mathbb{R}^d$
$\boldsymbol{\zeta}_j$ = "Nay" position $\in \mathbb{R}^d$
$\boldsymbol{\beta}_i$ = Ideal point $\in \mathbb{R}^d$

$$U_{i,j}(\text{Yeah}) = -\left\|\boldsymbol{\beta}_i - \boldsymbol{\psi}_j\right\|^2 + \epsilon_{i,j}$$
$$U_{i,j}(\text{Nay}) = -\left\|\boldsymbol{\beta}_i - \boldsymbol{\zeta}_j\right\|^2 + \nu_{i,j}$$

where $\nu_{i,j} - \epsilon_{i,j} \sim G_j$, and $y_{i,j} = 1 \Leftrightarrow U_{i,j}(\text{Yeah}) > U_{i,j}(\text{Nay})$,

$$\mu_j = \zeta_j^T \zeta_j - \psi_j^T \psi_j$$
$$\boldsymbol{\alpha}_j = 2(\boldsymbol{\psi}_j - \boldsymbol{\zeta}_j)$$

Policy Space representation



Binary factor models in general topological spaces

- Consider letting $\boldsymbol{\psi}_j, \boldsymbol{\zeta}_j, \boldsymbol{\beta}_i \in \mathcal{D}$, where $\mathcal{D}$ is a connected Riemannian manifold and define

$$U_{i,j}(\text{Yes}) = -\left\{d\left(\boldsymbol{\beta}_i, \boldsymbol{\psi}_j\right)\right\}^2 + \epsilon_{i,j},$$
$$U_{i,j}(\text{No}) = -\left\{d\left(\boldsymbol{\beta}_i, \boldsymbol{\zeta}_j\right)\right\}^2 + \nu_{i,j},$$

where $d\left(\boldsymbol{\beta}_i, \boldsymbol{\psi}_j\right)$ is the *geodesic distance* between $\boldsymbol{\beta}_i$ and $\boldsymbol{\psi}_j$ and $\nu_{i,j} - \epsilon_{i,j} \sim G_{\kappa_j}$.

- As before, $y_{i,j} = 1$ iff $U_{i,j}(\text{Yes}) > U_{i,j}(\text{No})$, so

$$\text{P}(y_{i,j} = 1 \mid \boldsymbol{\beta}_i, \zeta_j, \boldsymbol{\zeta}_j, \kappa_j) = G_{\kappa_j}\left(\left\{d\left(\boldsymbol{\beta}_i, \boldsymbol{\zeta}_j\right)\right\}^2 - \left\{d\left(\boldsymbol{\beta}_i, \boldsymbol{\psi}_j\right)\right\}^2\right).$$

Spherical factor models

- In the $\mathcal{S}^{K+1}$, the geodesic distance is given by $\rho_{K+1}(\boldsymbol{\psi}, \boldsymbol{\beta}) = \text{arccos}\left(\mathrm{x}_{\boldsymbol{\psi}}^T \mathrm{x}_{\boldsymbol{\beta}}\right)$, with, for example,

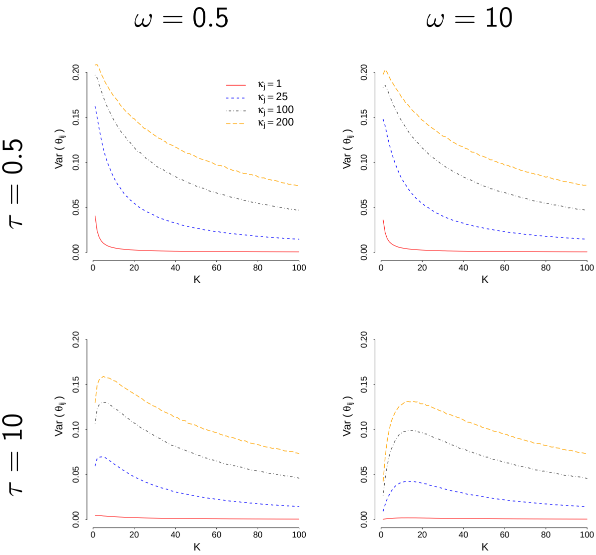
$$
\begin{aligned}
x_{\boldsymbol{\psi},1} &= \cos\psi_1 \cos\psi_2 \cos\psi_3 \cdots \cos\psi_{K-1}, \\
x_{\boldsymbol{\psi},2} &= \sin\psi_1 \cos\psi_2 \cos\psi_3 \cdots \cos\phi_{K-1}, \\
x_{\boldsymbol{\psi},3} &= \sin\psi_2 \cos\psi_3 \ldots \cos\psi_{K-1} \\
&\vdots \\
x_{\boldsymbol{\psi},K} &= \sin\psi_{K-1} \cos\psi_K, \\
x_{\boldsymbol{\psi},K+1} &= \sin\psi_K.
\end{aligned}
$$

- Yu, X., & Rodriguez, A. (2022). A Bayesian Approach to Spherical Factor Analysis for Binary Data. arXiv preprint arXiv:2008.05109.

Priors for spherical factor models

Standard von-Misses Fisher distributions on the sphere for $\{\psi_j\}_{j=1}^{J}$, $\{\zeta_j\}_{j=1}^{J}$ and $\{\beta_i\}_{i=1}^{I}$ will not work!

Variance of induced prior on $\theta_{i,j}$ - Von Misses-Fisher priors



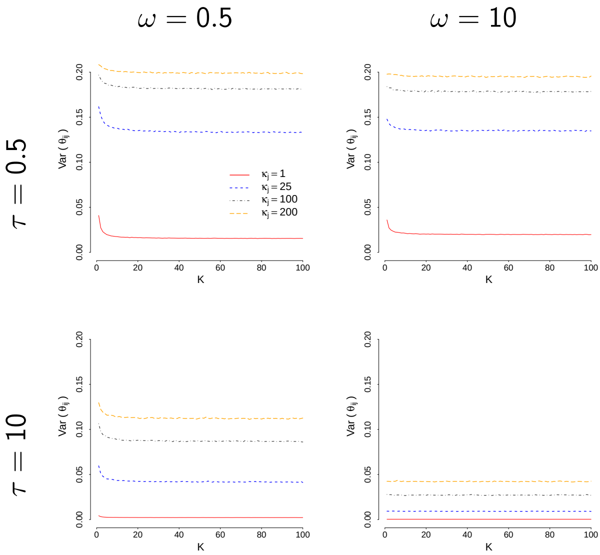
Spherical models

- We need a new class of priors on the sphere that allows for marginal variances of the angles to decrease with as new dimensions are added

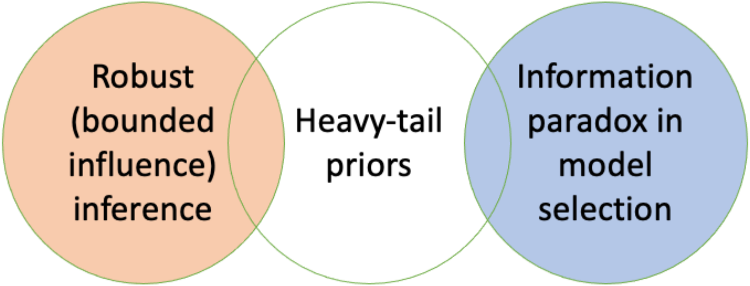
$$p(\boldsymbol{\phi} \mid \boldsymbol{\omega}) = \left(\frac{1}{2\pi}\right)^K 2^{K-1} \frac{1}{I_0(\omega_1)} \exp\left\{\omega_1 \cos \phi_1\right\} \prod_{k=2}^{K} \frac{1}{I_0(\omega_k)} \exp\left\{\omega_k \cos 2\phi_k\right\}$$

- Unlike the Euclidean case, we need the variance to decrease for both the ideal points and the Yes/No positions!

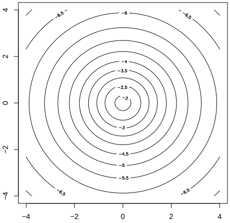
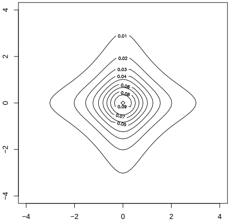
Variance of induced prior on $\theta_{i,j}$ - Von Misses-Fisher priors



The role of priors with heavy tails



Two philosophies ...

g-priors and its kin	Horseshoe and its kin
• Accounts for the “right” correlation among coefficients.	• Coefficients are independent a priori.
• “Non-directional”: Same tail behavior in every direction	• “Directional”: tails along axis are heavier than tails in other directons
	

Getting the best of both worlds:

- "Directional" g-priors:

$$\boldsymbol{\theta}_\gamma \mid \boldsymbol{\gamma}, \Lambda_\gamma, \sigma^2 \sim \mathrm{N}\left(0, \sigma^2 \Lambda_\gamma^{1/2} \left\{\mathrm{X}_\gamma^T \mathrm{X}_\gamma\right\}^{-1} \Lambda_\gamma^{1/2}\right)$$

with $\Lambda_\gamma = \mathrm{diag}\left\{\lambda_{\gamma,1}, \ldots \lambda_{\gamma,p_\gamma}\right\}$ and $\lambda_{\gamma,j} \sim H$.

- "Correlated" continuous shrinkage priors:

$$\boldsymbol{\theta} \mid \Lambda, \sigma^2 \sim \mathrm{N}\left(0, \sigma^2 \left\{\mathrm{X}^T \Lambda^{-1} \mathrm{X}\right\}^{-1}\right)$$

with $\Lambda = \mathrm{diag}\left\{\lambda_1, \ldots \lambda_p\right\}$ and $\lambda_j \sim H$.

Factor models

- A lot of the literature on continuous shrinkage priors has focused on making the Horseshoe a bit more flexible by making the distribution H more flexible by adding a couple of extra parameters.

- You could make the specification more flexible by setting a non-parametric prior on H (e.g., a Pòlya Tree centered on the half Cauchy distribution).

- Still somewhat speculative, this is work in progress!
  - Calibration?
  - How much can you really learn when you specify a non-parametric model further down in the hierarchy?
  - Efficient computation.

Concluding remarks

- A lot of the things that I learned from Pericchi 20 years ago still influence both my research and my teaching.
- I cannot believe it has been 20 years ...
- The school that he created in Venezuela starting in the id 80s and early 90s is still going strong, if in exile ...

Thank you!