Bayesian Approaches to Distribution Regression

Ho Chung Leon Law*, Dougal J. Sutherland*, Dino Sejdinovic, Seth Flaxman

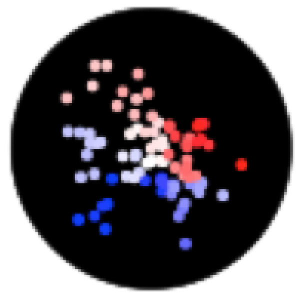
ho.law@spc.ox.ac.uk, dougals@gatsby.ucl.ac.uk, dino.sejdinovic@stats.ox.ac.uk, s.flaxman@imperial.ac.uk; arXiv:1705.04293

Distribution regression

Inputs are sample sets from distributions [e.g. 7]:

$$y_i = f^*(\mathbb{P}_i) + \varepsilon \qquad \mathbf{X}_i \overset{iid}{\sim} \mathbb{P}_i \qquad \hat{y}_i = f(\mathbf{X}_i)$$

Examples:

$\mathbb{P}_i$	$\mathbf{X}_i$	y_i
Multivariate distribution [7]	Sample from distribution	Entropy of a projection
Demographics of county population [2, 3]	AGEP SEX PINCP WKHP 45 2 14418.0 30.0 37 1 120152.0 40.0 63 1 11314.0 NaN 38 2 73092.0 40.0	30.3% Clinton 9.9% Trump 59.8% none/other
Galaxy velocities in a cluster [4, 5]		Cluster mass: $7 \times 10^{14} M_\odot$

Standard approach [e.g. 7]

- *Kernel mean embeddings*: kernel k, RKHS $\mathcal{H}$,
  $$\mu_{\mathbb{P}} = \mathbb{E}_{X\sim\mathbb{P}}[k(\cdot, X)] \in \mathcal{H}$$
- $\mu_{\mathbb{P}}$ fully characterizes $\mathbb{P}$ for many k
- Empirical mean estimator:
  $$\hat{\mu}_{\mathbb{P}} = \frac{1}{N}\sum_{j=1}^{N} k\left(\cdot, X^{(j)}\right)$$
- Ridge regression from $\hat{\mu}_{\mathbb{P}}$ to y_i:
  $$f(\hat{\mu}_{\text{test}}) = \sum_{i=1}^{n} \alpha_i \langle \hat{\mu}_{\text{test}}, \hat{\mu}_i\rangle_{\mathcal{H}} = \sum_{i=1}^{n}\alpha_i \sum_{j=1}^{N_{\text{test}}}\sum_{j'=1}^{N_i} k\left(X_j^{\text{test}}, X_{j'}^{i}\right)$$
- We make a landmark approximation (RBF net):
  $$f(\hat{\mu}) = \beta^\mathsf{T}\hat{\mu}(\mathbf{u}) = \sum_{\ell=1}^{s}\beta_\ell\,\hat{\mu}(u_\ell)$$

Sources of uncertainty

- Uncertainty about regression weights α
- Mean embeddings μ_i not seen exactly
  - Should "trust" points with small N_i less

Bayesian linear regression

- Standard BLR model (similar to e.g. [3])
  - Handles regression weight β uncertainty
  - Assumes $\hat{\mu}_i$ known exactly
- Normal prior over regression weights $\beta \sim \mathcal{N}(0, \rho^2)$
- Observations $y_i \mid \mathbf{X}_i, \beta \sim \mathcal{N}(\beta^\mathsf{T}\hat{\mu}_i(\mathbf{u}), \sigma^2)$
- Gives normal $y_i \mid \mathbf{X}_i$ with hyperparameters σ, ρ

Shrinkage

- New model
  - Handles $\hat{\mu}_i$ uncertainty
  - Point estimate for β to retain conjugacy
- Uses Bayesian nonparametric model for $\hat{\mu}_i$ [1]:
  - Prior: $\mu_i \sim \mathcal{GP}(m_0, \eta r(\cdot,\cdot))$
  - Likelihood: "observed" at points $\mathbf{u}$, CLT:
    $$\hat{\mu}_i(\mathbf{u}) \mid \mu_i(\mathbf{u}) \sim \mathcal{N}(\mu_i(\mathbf{u}), \Sigma_i)$$
  - Closed-form GP posterior for $\hat{\mu}_i \mid \mathbf{X}_i$
  - Similar to James-Stein shrinkage
- Observations $y_i \mid \mu_i, f \sim \mathcal{N}(\langle f, \mu_i\rangle_{\mathcal{H}}, \sigma^2)$
- Say $f(\cdot) = \sum_{\ell=1}^{s}\alpha_\ell k(\cdot, z_\ell)$ (representer theorem)
- Predictive: $y_i \mid \mathbf{X}_i, \alpha, \mathbf{z} \sim \mathcal{N}(\xi_i^\alpha, \nu_i^\alpha)$
  $$\xi_i^\alpha = \alpha^\top R_{\mathbf{z}}\left(R + \frac{\Sigma_i}{N_i}\right)^{-1}(\hat{\mu}_i - m_0) + \alpha^\mathsf{T} m_0$$
  $$\nu_i^\alpha = \alpha^\mathsf{T}\left(R_{\mathbf{zz}} - R_{\mathbf{z}}\left(R + \frac{\Sigma_i}{N_i}\right)^{-1} R_{\mathbf{z}}^\mathsf{T}\right)\alpha + \sigma^2$$
- MAP estimate for α, σ, maybe $\mathbf{z}$, kernel params…
- Optimise with backprop (TensorFlow)

Bayesian Distribution Regression

- Full uncertainty model (BDR)
  - Shrinkage posterior for $\hat{\mu}_i$
  - BLR-like posterior for weights α
- Not conjugate
- MCMC for inference about α (Stan)

Results on synthetic data

- Toy problem to examine input uncertainty:
  - Labels y_i uniform over $[4, 8]$
  - 5d data points: $\left[x_j^i\right]_\ell \mid y_i \overset{iid}{\sim} \frac{1}{y_i}\Gamma\left(\frac{y_i}{2}, \frac{1}{2}\right)$
  - $(s_5, 25, 25, 100 - s_5)$% have $N_i = (5, 20, 100, 1000)$
- BDR ≈ shrinkage < BLR in NLL, MSE:

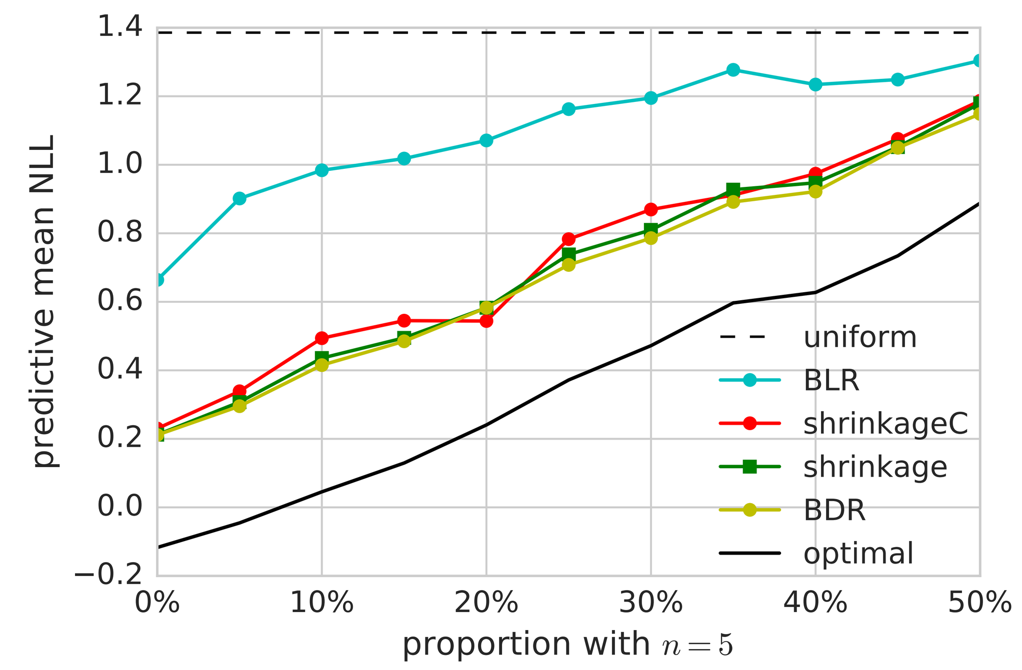


Figure: Negative log predictive likelihoods.

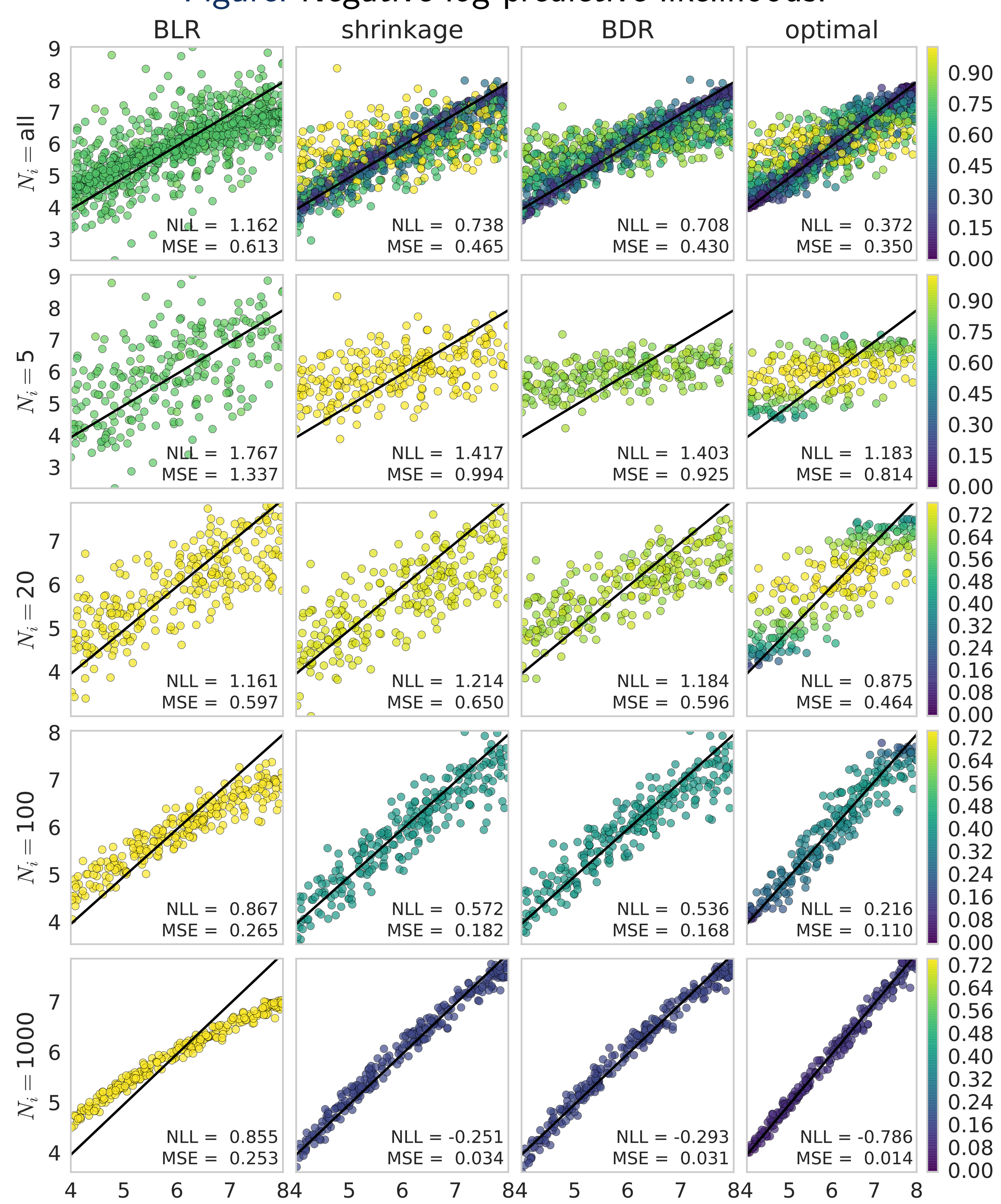
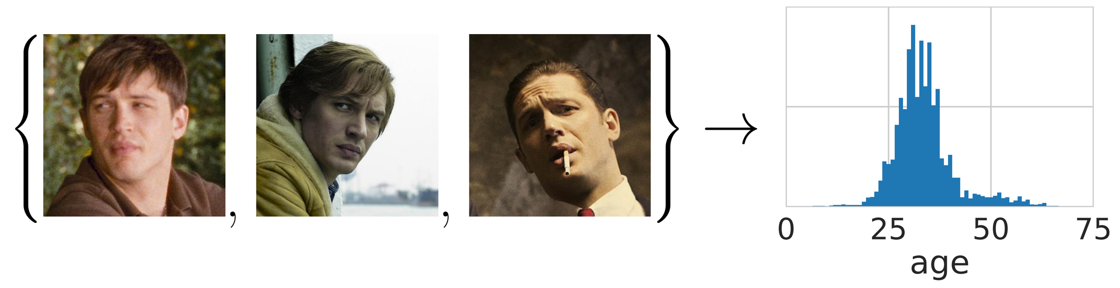


Figure: Bags are dots: horizontal position is true y_i, vertical is predictive mean, colour is predictive std. $s_5 = 25$.

- Similar experiment with constant N_i, added noise:
  - BDR ≈ BLR < shrinkage in NLL, MSE
- BDR able to take advantage of both settings

Age prediction from images



- IMDb database from [6]
  - *Very* noisy labels in the dataset
- Group pictures of actors, predict mean age
- Features: last hidden layer from [6]'s CNN
- Lots of variation in N_i:

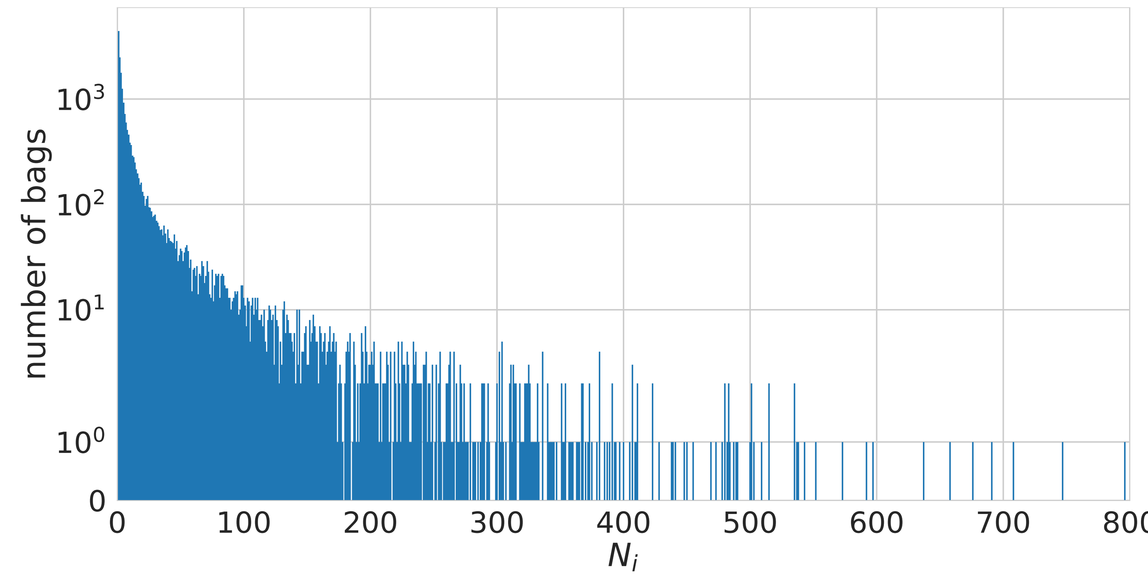


Figure: Histogram of N_i. 22% have $N_i = 1$.

- Shrinkage really helps!

Method	NLL	RMSE
CNN	3.80 (0.03)	10.25 (0.22)
RBF network	–	9.51 (0.20)
BLR	3.68 (0.02)	9.55 (0.19)
shrinkage	**3.54** (0.02)	**9.28** (0.20)

Table: Results across ten data splits (means and standard deviations). Here **shrinkage** performs the best across all 10 runs in both metrics. CNN takes the mean of the predictive distributions of [6].

References

[1] S. Flaxman, D. Sejdinovic, J. Cunningham and S. Filippi. 'Bayesian Learning of Kernel Embeddings'. In: *UAI*. 2016. arXiv: 1603.02160.

[2] S. Flaxman, D. J. Sutherland, Y.-X. Wang and Y.-W. Teh. *Understanding the 2016 US Presidential Election using ecological inference and distribution regression with census microdata*. 2016. arXiv: 1611.03787.

[3] S. Flaxman, Y.-X. Wang and A. J. Smola. 'Who Supported Obama in 2012?: Ecological inference through distribution regression'. In: *KDD*. ACM. 2015.

[4] M. Ntampaka, H. Trac, D. J. Sutherland, N. Battaglia, B. Póczos and J. Schneider. 'A Machine Learning Approach for Dynamical Mass Measurements of Galaxy Clusters'. In: *The Astrophysical Journal* 803.2 (2015), p. 50. arXiv: 1410.0686.

[5] M. Ntampaka, H. Trac, D. J. Sutherland, S. Fromenteau, B. Poczos and J. Schneider. 'Dynamical Mass Measurements of Contaminated Galaxy Clusters Using Machine Learning'. In: *The Astrophysical Journal* 831.2 (2016), p. 135. arXiv: 1509.05409.

[6] R. Rothe, R. Timofte and L. V. Gool. 'Deep expectation of real and apparent age from a single image without facial landmarks'. In: *International Journal of Computer Vision (IJCV)* (July 2016).

[7] Z. Szábo, B. K. Sriperumbudur, B. Póczos and A. Gretton. 'Learning Theory for Distribution Regression'. In: *JMLR* 17.152 (2016), pp. 1–40. arXiv: 1411.2066.