Bayesian Approaches to Distribution Regression

Ho Chung Leon Law[*]
University of Oxford
ho.law@spc.ox.ac.uk

Dougal J. Sutherland[*]
University College London
dougal@gmail.com

Dino Sejdinovic
University of Oxford
dino.sejdinovic@stats.ox.ac.uk

Seth Flaxman
Imperial College London
s.flaxman@imperial.ac.uk

Abstract

Distribution regression has recently attracted much interest as a generic solution to the problem of supervised learning where labels are available at the group level, rather than at the individual level. Current approaches, however, do not propagate the uncertainty in observations due to sampling variability in the groups. This effectively assumes that small and large groups are estimated equally well, and should have equal weight in the final regression. We construct a Bayesian distribution regression formalism that accounts for this uncertainty, improving the robustness and performance of the model when group sizes vary. We can obtain MAP estimates for some models with backpropagation, while the full propagation of uncertainty requires MCMC-based inference. We demonstrate our approach on an illustrative toy dataset as well as a challenging age prediction problem.

1 Introduction

Distribution regression is the problem of learning a regression function from samples of a distribution to a single set-level label. For example, we might infer the sentiment of sentences or paragraphs based on word features, predict the label of an image based on small patches, or even perform traditional parametric statistical inference by learning a function from sets of samples to the parameter values. Recent years have seen many wide-ranging applications of this framework, including inferring summary statistics in Approximate Bayesian Computation [10], estimating Expectation Propagation messages [7], predicting the aggregate voting behaviour of demographic groups [3, 5], and learning the total mass of dark matter halos from observable galaxy velocities [13, 14].

One appealing approach to the distribution regression problem [11, 20, 21, 3, 7, 9, 10] is to represent the input set of samples by their kernel mean embedding, a point in a reproducing kernel Hilbert space, and then apply standard kernel methods. In this framework, however, each distribution is simply represented by its empirical mean embedding, ignoring that large sample sets are understood much more precisely than small ones. Most studies also use point estimates for the regression function.

We propose a set of Bayesian approaches to distribution regression. We build on the recently proposed Bayesian nonparametric prior over kernel mean embeddings [4] to account for uncertainty in the kernel mean embeddings, and then use a sparse representation of the desired function in the RKHS for prediction in the regression model. For this model, we use MAP estimation of the non-conjugate parameters. Bayesian linear regression instead accounts for uncertainty in the regression model. Finally, we can combine the treatment of the two sources of uncertainty into a fully Bayesian model, combining both source of uncertainty, and use Hamiltonian Monte Carlo for efficient inference. Depending on the setting, each approach may be useful.

[*] These authors contributed equally.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

This short paper gives a necessarily abbreviated account. For a more complete treatment, we encourage the interested reader to consult the full version at `arxiv.org/abs/1705.04293`.

2 Background

2.1 Problem overview

In distribution regression, we wish to map probability distributions to labels. The challenge of distribution regression goes beyond the standard supervised learning setting: we do not have access to exact input-output pairs since the true inputs, complex probability distributions, are observed only through samples from that distribution. Our observations are structured as:

$$\left(\{x_j^1\}_{j=1}^{N_1}, y_1\right), \ldots, \left(\{x_j^n\}_{j=1}^{N_n}, y_n\right), \tag{1}$$

so that each bag $\{x_j^i\}_{j=1}^{N_i}$ has a label y_i along with N_i individual observations $x_j^i \in \mathcal{X}$. We assume that the observations $\{x_j^i\}_{j=1}^{N_i}$ are i.i.d. samples from some unobserved distribution P_i, and that the true label y_i depends only on P_i. We wish to avoid making strong parametric assumptions on P_i. We assume the labels y_i are real-valued; the full paper shows an extension to binary classification.

The standard approach to distribution regression relies on kernel mean embeddings and kernel ridge regression. We assume we have a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, whose corresponding reproducing kernel Hilbert space (RKHS) we call $\mathcal{H}_k$. The kernel mean embedding of a probability measure P on $\mathcal{X}$, which exists at least when k is bounded, is

$$\mu_{\mathsf{P}} = \int k(\cdot, x)\, \mathsf{P}(dx) \in \mathcal{H}_k. \tag{2}$$

Notice that μ_{P} serves as a (likely infinite-dimensional) vectorial representation of P. For so-called *characteristic* kernels [18], every probability measure has a unique embedding.

2.2 Estimating mean embeddings

For a set of samples $\{x_j\}_{j=1}^n$ drawn iid from P, the empirical estimator of μ_{P} is given by

$$\widehat{\mu_{\mathsf{P}}} = \mu_{\widehat{\mathsf{P}}} = \int k(\cdot, x)\, \hat{\mathsf{P}}(dx) = \frac{1}{n}\sum_{j=1}^{n} k(\cdot, x_j). \tag{3}$$

This is the standard estimator used by previous distribution regression approaches. But (3) is an empirical mean estimator in a high- or infinite-dimensional space, and is thus subject to the well-known *Stein phenomenon*, so that its performance is dominated by the James-Stein shrinkage estimators. Indeed, Muandet et al. [12] studied shrinkage estimators for mean embeddings, which can substantially improve performance in some settings [16].

Flaxman et al. [4] proposed a Bayesian analogue of shrinkage estimators, which we now review. This approach consists of (1) a Gaussian Process prior $\mu_{\mathsf{P}} \sim \mathcal{GP}(m_0, r(\cdot, \cdot))$ on $\mathcal{H}_k$, where r is selected to ensure that $\mu_{\mathsf{P}} \in \mathcal{H}_k$ almost surely[1] and (2) a normal likelihood $\widehat{\mu_{\mathsf{P}}}(\mathbf{x}) \mid \mu_{\mathsf{P}}(\mathbf{x}) \sim \mathcal{N}(\mu_{\mathsf{P}}(\mathbf{x}), \Sigma)$. Conjugacy of the prior and the likelihood leads to the Gaussian process posterior on the true embedding μ_{P} given the "observed" empirical embedding $\widehat{\mu_{\mathsf{P}}}$ at a given set of locations $\mathbf{x}$ where the embeddings are evaluated; see (4). The posterior mean is then essentially identical to a particular shrinkage estimator of [12], but we also gain a closed form uncertainty estimate.

This model accounts for the uncertainty in the number of samples N_i, shrinking the embeddings for small sample sizes more. We will see this is essential in the context of distribution regression, particularly when training set sizes are imbalanced.

2.3 Standard approaches to distribution regression

Following Szábo et al. [20], assume that the probability distributions P_i are each drawn randomly from some unknown meta-distribution over probability distributions, and take a two-stage approach:

[1] For our Gaussian kernel, we can either choose $r = k$, which *almost* gives this property, or choose r as a convolution of k; see the full paper for details.

we first use the empirical kernel mean estimator (3) to separately estimate the mean of each group. Next, we use kernel ridge regression [17] to learn a function f:

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{H}_K} \sum_i (y_i - f(\widehat{\mu_i}))^2 + \lambda \|f\|^2_{\mathcal{H}_K},$$

where K represents a second-level kernel $K : \mathcal{H}_k \times \mathcal{H}_k \to \mathbb{R}$. This can be simply implemented using the kernel trick [11]. For even modestly-sized datasets, however, this can be quite expensive: the kernel matrix over distributions has $\mathcal{O}(n^2)$ entries, but entry (i, j) takes time $\mathcal{O}(N_i N_j)$ to compute. Many applications have thus approximated $\mathcal{H}_k$ with random Fourier features [15].

We take a simpler approach here and use landmark points drawn randomly from the observations, effectively yielding radial basis networks [2] with a mean pooling operation. Specifically, our base model is the following: we select landmark points $\mathbf{u} = \{u_\ell\}_{\ell=1}^d$. Each point $x_j^i \in \mathbb{R}^p$ is mapped to

$$\phi(x_j^i) = [k(x_j^i, u_1), \ldots, k(x_j^i, u_d)]^\top \in \mathbb{R}^d.$$

We then estimate the mean embedding $\hat{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \phi(x_j^i)$ for each bag in a minibatch, and then obtain real-valued labels as $\hat{y}_i = \beta^\mathsf{T} \hat{\mu}_i + b$ for regression weights β and intercept b. We use mean squared error as the loss function, and learn with the Adam optimizer [8]. We regularise with early stopping on a validation set, as well as an explicit L_2 penalty corresponding to a normal prior on β.

3 Bayesian models

We propose three different Bayesian models, with each model encoding different types of uncertainty.

3.1 Mean shrinkage pooling model

A shortcoming of the standard approach is that it ignores uncertainty in the first level of estimation due to varying number of samples in each bag. Ideally we would estimate not just the mean embedding per bag, but also a measure of the sample variance, in order to propagate this information regarding uncertainty from the bag size through. Bayesian tools provide a natural framework for this problem.

We can use the Bayesian nonparametric prior over kernel mean embeddings [4] described in Section 2.2, and 'observe' the empirical embeddings at the landmark points $\mathbf{u_i}$ (chosen at random from the dataset, or via k-means). Using a Gaussian process prior $\mu_i \sim \mathcal{GP}(m_0, \eta r(\cdot, \cdot))$ and a covariance of Σ in the likelihood gives us a closed form posterior Gaussian process, whose evaluation at points $\mathbf{h} = \{h_s\}_{s=1}^k$ is:

$$\mu_i(\mathbf{h}) \mid \mathbf{x_i} \sim \mathcal{N}\left(R_\mathbf{h} \left(R + \Sigma_i/N_i\right)^{-1} (\hat{\mu}_i - m_0) + m_0, R_\mathbf{hh} - R_\mathbf{h} \left(R + \Sigma_i/N_i\right)^{-1} R_\mathbf{h}^\top\right) \quad (4)$$

where $R_{st} = \eta r(u_s, u_t), (R_\mathbf{hh})_{st} = \eta r(h_s, h_t), (R_\mathbf{h})_{st} = \eta r(h_s, u_t)$, and $\mathbf{x_i}$ denotes the set $\{x_j^i\}_{j=1}^{N_i}$. We take the prior mean m_0 to be the mean of $\hat{\mu}_i$; when K is linear, this corresponds to shrinking predictions towards the mean prediction. Smaller η correspond to stronger shrinkage towards m_0. We take Σ to be the mean of the empirical covariances of $\{\varphi(x_j^i)\}_{j=1}^{N_i}$.

Motivated by the representer theorem, we use a regression function of the form $f = \sum_{s=1}^{n_z} \alpha_s k(\cdot, z_s)$, with each $z_s \in \mathbb{R}^d$ a landmark point. Thus $y_i \mid \mu_i, \alpha \sim \mathcal{N}\left(\alpha^\top \mu_i(\mathbf{z}), \sigma^2\right)$, where $\mu_i(\mathbf{z}) = [\mu_i(z_1), \ldots, \mu_i(z_s)]^\top$. For fixed α, the predictive distribution (taking $m_0 = 0$ for simplicity) is:

$$y_i \mid \mathbf{x}_i, \alpha \sim \mathcal{N}\left(\alpha^\mathsf{T} R_\mathbf{z} \left(R + \Sigma_i/N_i\right)^{-1} \hat{\mu}_i, \alpha^\mathsf{T} \left(R_\mathbf{zz} - R_\mathbf{z} \left(R + \Sigma_j/N_j\right)^{-1} R_\mathbf{z}^\mathsf{T}\right) \alpha + \sigma^2\right).$$

Taking a prior $\alpha \sim \mathcal{N}(0, \rho^2 K_\mathbf{z}^{-1})$, we can easily learn a MAP estimate for α, σ, η, and potentially $\mathbf{z}$ or any parameters of k via backpropagation, while maintaining the full account of uncertainty for μ_i.

3.2 Bayesian linear regression model

An alternative approach is to encode uncertainty over the regression parameters β only:

$$\beta \sim \mathcal{N}(0, \rho^2 I) \qquad y_i \mid \mathbf{x}_i, \beta \sim \mathcal{N}(\beta^\mathsf{T} \hat{\mu}_i, \sigma^2),$$

obtaining Bayesian linear regression over empirical mean embeddings. Here we work directly with our finite-dimensional approximation $\hat{\mu}_i$. We can easily find $y_i \mid \mathbf{x}_i$, and use backpropagation on the marginal log-likelihood [see e.g. 1] to learn σ, ρ, and any kernel parameters. This model provides uncertainty over the regression function, but ignores uncertainty in mean embeddings.

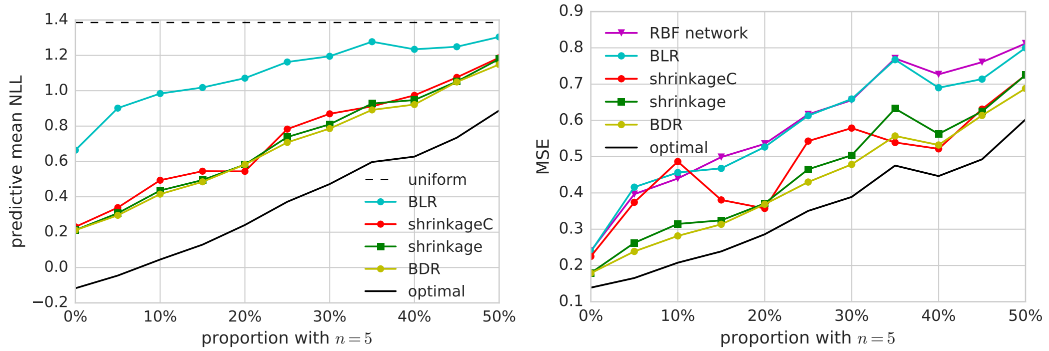

Figure 1: Results for the experiment of Section 4: predictive mean negative log-likelihoods at left, mean squared error at right. shrinkage and shrinkageC refer to the method of Section 3.1 with $r = k$ and the convolutional r, respectively; BLR the method of Section 3.2; BDR that of Section 3.3. Bayes-optimal results also shown for context. The best constant predictor achieved MSE about 1.3.

3.3 Bayesian distribution regression

From a modelling perspective, it is natural to combine the two Bayesian approaches above, fully propagating uncertainty in estimation of the mean embedding and of the regression coefficients α. Unfortunately, conjugate Bayesian inference is no longer available. Thus, we consider a Markov chain Monte Carlo (MCMC) sampling based approach, using Hamiltonian Monte Carlo (HMC) for efficient inference. Whereas inference above used gradient descent to maximise the marginal likelihood, here we use automatic differentiation to calculate the gradient of the joint log-likelihood and follow this gradient as we perform sampling over the parameters we wish to infer. We can still exploit the conjugacy of the mean shrinkage layer, obtaining closed form expressions for the posterior over the mean embeddings. Conditional on the mean embeddings, we have a Bayesian linear regression model with parameters β which we sample with HMC, specifically NUTS [6, 19].

4 Numerical experiments

We consider the following toy problem:

$$y_j \sim \text{Uniform}(4, 8) \qquad \left[x_j^i\right]_\ell \mid y_j \overset{i.i.d.}{\sim} \frac{1}{y_j}\left[\Gamma\left(\frac{y_j}{2}, \frac{1}{2}\right)\right] \text{ with } \ell = 1, \ldots, 5.$$

Each dataset has 25% bags with $N_i = 20$, and 25% with $N_i = 100$; the remainder have some portion with $N_i = 5$ and the remainder with $N_i = 1000$. Figure 1 shows predictive negative log-likelihood and mean squared error results for the various models, as well as the performance of the Bayes-optimal predictor and the best data-independent predictor for context. We can see that shrinkage and the full-Bayesian model significantly outperform BLR and the baseline model, both in predictive likelihoods and in mean squared error.

The long version of the paper also demonstrates a case where, with larger bag sizes and adding additional noise to the problem, Bayesian linear regression outperforms shrinkage. The full-uncertainty model still performs best.

In the full paper, we also consider a real problem where we observe multiple images of a single person and attempt to predict their mean age. Using a deep kernel defined by a pretrained neural network, we see in this case that shrinkage, and the full uncertainty model, yield better predictions and better uncertainty estimates than baselines.

5 Conclusion

We have provided a method for accounting for uncertainty in the observation of distributions within distribution regression methods. We expect that powerful future distribution regression approaches will need to incorporate this aspect of uncertainty, and that our methods provide a strong and generic building block for doing so.

References

[1] C.M. Bishop. *Pattern recognition and machine learning*. Springer New York, 2006.

[2] David S Broomhead and David Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, DTIC Document, 1988.

[3] Seth Flaxman, Yu-Xiang Wang, and Alexander J Smola. Who supported Obama in 2012?: Ecological inference through distribution regression. In *KDD*, pages 289–298. ACM, 2015.

[4] Seth Flaxman, Dino Sejdinovic, John P. Cunningham, and Sarah Filippi. Bayesian learning of kernel embeddings. In *UAI*, 2016.

[5] Seth Flaxman, Dougal J. Sutherland, Yu-Xiang Wang, and Yee-Whye Teh. Understanding the 2016 US presidential election using ecological inference and distribution regression with census microdata. 2016. arXiv:1611.03787.

[6] Matthew D. Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *JMLR*, pages 1593–1623, 2014.

[7] Wittawat Jitkrittum, Arthur Gretton, Nicolas Heess, S. M. Ali Eslami, Balaji Lakshminarayanan, Dino Sejdinovic, and Zoltán Szabó. Kernel-Based Just-In-Time Learning for Passing Expectation Propagation Messages. In *UAI*, 2015.

[8] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. arXiv:1412.6980.

[9] David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya Tolstikhin. Towards a learning theory of cause-effect inference. In *ICML*, 2015.

[10] J. Mitrovic, D. Sejdinovic, and Y.W. Teh. DR-ABC: Approximate Bayesian Computation with Kernel-Based Distribution Regression. In *ICML*, pages 1482–1491, 2016.

[11] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *NIPS*, 2012. arXiv:1202.6504.

[12] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Arthur Gretton, and Bernhard Schoelkopf. Kernel mean estimation and stein effect. In *ICML*, 2014.

[13] Michelle Ntampaka, Hy Trac, Dougal J. Sutherland, Nicholas Battaglia, Barnabás Póczos, and Jeff Schneider. A machine learning approach for dynamical mass measurements of galaxy clusters. *The Astrophysical Journal*, 803(2):50, 2015. ISSN 1538-4357. arXiv:1410.0686.

[14] Michelle Ntampaka, Hy Trac, Dougal J. Sutherland, S. Fromenteau, B. Poczos, and Jeff Schneider. Dynamical mass measurements of contaminated galaxy clusters using machine learning. *The Astrophysical Journal*, 831(2):135, 2016. arXiv:1509.05409.

[15] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *NIPS*, pages 1177–1184, 2007.

[16] Aaditya Ramdas and Leila Wehbe. Nonparametric independence testing for small sample sizes. In *IJCAI*, 2015. arXiv:1406.1922.

[17] Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. In *ICML*, 1998.

[18] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *JMLR*, 99:1517–1561, 2010.

[19] Stan Development Team. Stan: A c++ library for probability and sampling, version 2.5.0, 2014. URL `http://mc-stan.org/`.

[20] Zoltán Szábo, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Leraning theory for distribution regression. *JMLR*, 17(152):1–40, 2016. arXiv:1411.2066.

[21] Yuya Yoshikawa, Tomoharu Iwata, and Hiroshi Sawada. Latent support measure machines for bag-of-words data classification. In *NIPS*, pages 1961–1969, 2014.