Kernel Distribution Embeddings

Classical kernel methods have addressed the mapping of individual points to feature space. A generalization is towards feature space representations of probability distributions. We collectively refer to such mappings as distribution embeddings.

A natural application of kernel distribution embeddings is in testing for similarities between samples from probability distributions. We refer to the distance between two distribution embeddings as the *maximum mean discrepancy* (MMD). We have formulated a two-sample test [1] (of whether two distributions are the same), and showed the independence test in[1] (of whether two random variables observed together are statistically independent) is a special case. A further application of the MMD as independence criterion is in feature selection, where we maximize dependence between features and labels [3]. We have further developed alternative independence tests based on space partitioning approaches and classical divergence measures (such as the ℓ_1 distance and KL-divergence) [7, 2].

A recent application uses kernel means in visualization. When using a power-of-cosine kernel for distributions on the projective sphere, the kernel mean can be represented as a symmetric tensor. In the context of diffusion MRI, this permits an efficient visual and quantitative analysis of the uncertainty in nerve fiber estimates, which can inform the choice of MR acquisition schemes and mathematical models.[2]

We have also used the kernel means embedding to develop a variant of an SVM which operates on distributions rather than points [11], permitting modeling of input uncertainties. One can prove a generalized representer theorem for this case, and in the special case of Gaussian input uncertainties and Gaussian kernel SVMs, it leads to a multi-scale SVM, akin to an RBF network with variable widths, which is still trained by solving a quadratic optimization problem.

Given that the MMD depends on the particular kernel chosen, we proposed two kernel selection strategies [15, 8], the earlier one relying on a classification interpretation of the MMD, and the later one explicitly minimizing the probability of Type II error of the associated two-sample test (that is, the probability of wrongly accepting that two unlike distributions are the same, given samples from each).

A natural question to consider is whether the MMD constitutes a metric on distributions, and is zero if and only if the distributions are the same. When this holds, the RKHS is said to be *characteristic*. We have determined straightforward necessary and sufficient conditions on translation invariant kernels for injectivity, for distributions on compact and non-compact subsets of $\mathbb{R}^d$ [5]: specifically, the Fourier transform of the kernel should be supported on all of $\mathbb{R}^d$. Gaussian, Laplace, and B-spline kernels satisfy this requirement. The MMD is a member of a larger class of metrics on distributions, known as the integral probability metrics. In [16, 4], we provide estimates of integral probability metrics on $\mathbb{R}^d$ which are taken over function classes that are not RKHSs, namely the Wasserstein distance (functions in the unit Lipschitz semi-norm ball) and the Dudley metric (functions in the unit bounded Lipschitz norm ball), and establish strong consistency of our estimators. Comparing the MMD and these two distances, the MMD converges fastest, and at a rate independent of the dimensionality d of the random variables - by contrast, rates for the classical Wasserstein and Dudley metrics worsen when d grows.

Embeddings of distributions can be generalized to yield embeddings of conditional distributions. A first application of conditional distribution embeddings is to Bayesian inference on graphical models. We have developed two approaches: in the first [14, 13], the messages are conditional density functions, subject to smoothness constraints; these were orders of magnitude faster than competing nonparametric BP approaches, yet more accurate, on problems including depth reconstruction from 2-D images and robot orientation recovery. In the second approach [6], the distributions are represented directly as embeddings of conditional distributions in RKHSs, allowing greater generality (for instance, one can define distributions over structured objects such as strings or graphs, for which probability densities may not exist). We showed the conditional mean embedding to be a solution to a vector valued regression problem [9], which allows us to formulate sparse estimates.

[1] Gretton et al. *A Kernel Statistical Test of Independence*. NIPS 2007

[2] T. Schultz, L. Schlaffke, B. Schölkopf, T. Schmidt-Wilcke. HiFiVE: A Hilbert Space Embedding of Fiber Variability Estimates for Uncertainty Modeling and Visualization. Submitted to: Eurographics Conference on Visualization (EuroVis), 2013

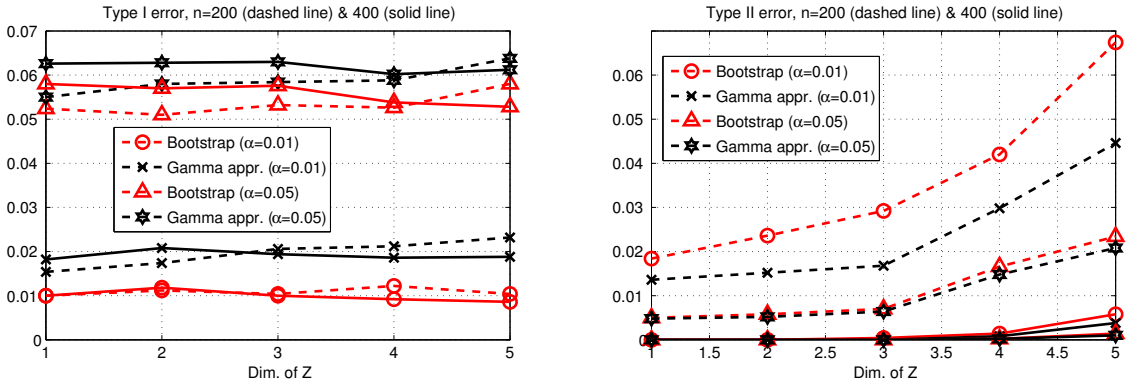

Figure 0.2: Type I (left) and Type II (right) errors of KCI-test in a simulated situation.

A second application of conditional distribution embeddings is to reinforcement learning. In [10], we estimate the optimal value function for a Markov decision process using conditional distribution embeddings, and the associated policy. An illustration is given in the first figure set, where the value function is estimated in two rooms connected by a corridor, where the agent has access only to images of the wall textures. This work was generalized to partially observable Markov decision processes in [12], where the kernel Bayes' law was used to integrate over distributions of the hidden states.

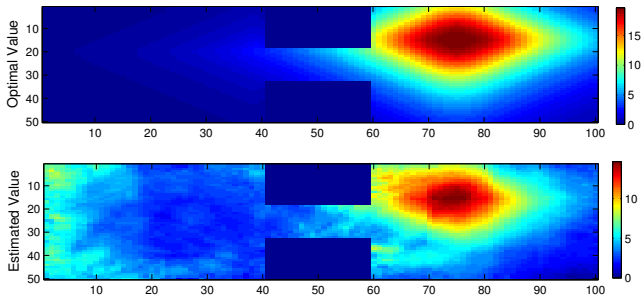

Figure 0.1: Top: Optimal value function, two rooms dataset. Bottom: Value function estimated via kernel MDP.

A final application of conditional mean embeddings is in testing for conditional independence (CI). Generally speaking, the CI between X and Y given Z, denoted by $X \perp\!\!\!\perp Y|Z$, allows us to drop Y when constructing a probabilistic model for X with (Y, Z), resulting in a parsimonious representation.

Testing for CI is much more difficult than that for unconditional independence. For CI tests, traditionally, one either focuses on the discrete case, or imposes simplifying assumptions to deal with the continuous case – in particular, the variables are often assumed to have linear relations with additive Gaussian errors. In that case, CI can be easily tested. However, nonlinearity and non-Gaussian noise are frequently encountered in practice, and hence the linear-Gaussian assumption can lead to incorrect conclusions.

Recently, practical methods have been proposed for testing CI for continuous variables without assuming a functional form between the variables as well as the data distributions, which is the case we are concerned with. Existing methods are based on explicit estimation of the conditional densities or discretization of the conditioning set Z, or exploit bootstrap to determine the rejection region. These methods require a large sample size and tend to be unreliable when the number of conditioning variables increases.

We proposed a Kernel-based Conditional Independence test (KCI-test [17]) which avoids the above drawbacks. In particular, based on appropriate characterizations of the CI relationship $X \perp\!\!\!\perp Y|Z$ based on conditional cross-covariance operators, we define a simple test statistic which can be easily calculated from the kernel matrices associated with X, Y, and Z; most importantly, we further derive its asymptotic distribution under the null hypothesis, and provide ways to estimate such a distribution. Finally CI can be tested conveniently. In this procedure we do not explicitly estimate the conditional or joint densities, nor discretize the conditioning variables. Our method is computationally appealing and is less sensitive to the dimensionality of Z compared to other methods. Our results generalize previous results on unconditional independence testing[1] as a special case. This is the first time that the null distribution of the kernel-based statistic for CI testing has been derived.

Publications

Journal Articles

[1] A Gretton, K Borgwardt, M Rasch, B Schölkopf, and A Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 3 2012. 1

[2] A Gretton and L Györfi. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010. 1

[3] L Song, A Smola, A Gretton, J Bedo, and K Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13:1393–1434, 5 2012. 1

[4] B Sriperumbudur, K Fukumizu, A Gretton, B Schölkopf, and G Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012. 1

[5] BK Sriperumbudur, A Gretton, K Fukumizu, B Schölkopf, and GRG Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 4 2010. 1

Articles in Conference Proceedings

[6] K Fukumizu, L Song, and A Gretton. Kernel Bayes' rule. In J Shawe-Taylor, RS Zemel, P Bartlett, F Pereira, and KQ Weinberger, editors, *25th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1737–1745, Granada, Spain, 2011. Curran Associates, Inc. 1

[7] A Gretton and L Györfi. Nonparametric independence tests: Space partitioning and kernel approaches. In Y Freund, L Györfi, and G Turán T Zeugmann, editors, *19th International Conference on Algorithmic Learning Theory (ALT08)*, pages 183–198, Budapest, Hungary, 10 2008. Springer. 1

[8] A Gretton, B Sriperumbudur, D Sejdinovic, H Strathmann, S Balakrishnan, M Pontil, and K Fukumizu. Optimal kernel choice for large-scale two-sample tests. In P Bartlett, FCN Pereira, CJC Burges, L Bottou, and KQ Weinberger, editors, *26th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1214–1222, Lake Tahoe, Nevada, USA, 2012. 1

[9] S Grünewälder, G Lever, L Baldassarre, S Patterson, A Gretton, and M Pontil. Conditional mean embeddings as regressors. In J Langford and J Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1823–1830, Edinburgh, Scotland, GB, 2012. Omnipress. 1

[10] S Grünewälder, G Lever, Ll Baldassarre, M Pontil, and A Gretton. Modelling transition dynamics in mdps with RKHS embeddings. In J Langford and J Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning*, pages 535–542, Edinburgh, Scotland, GB, 2012. Omnipress. 2

[11] K Muandet, K Fukumizu, F Dinuzzo, and B Schölkopf. Learning from distributions via support measure machines. In P Bartlett, FCN Pereira, CJC Burges, L Bottou, and KQ Weinberger, editors, *26th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 10–18, Lake Tahoe, Nevada, USA, 2012. 1

[12] Y Nishiyama, A Boularias, A Gretton, and K Fukumizu. Hilbert space embeddings of POMDPs. In *Conference on Uncertainty in Artificial Intelligence (UAI 2012)*, Catalina Island, USA, 2012. 2

[13] L Song, A Gretton, D Bickson, Y Low, and C Guestrin. Kernel belief propagation. In G Gordon, D Dunson, and M Dudík, editors, *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 707–715, Fort Lauderdale, FL, USA, 2011. JMLR. 1

[14] L Song, A Gretton, and C Guestrin. Nonparametric tree graphical models. In YW Teh and M Titterington, editors, *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 765–772, Sardinia, Italy, 2010. JMLR. 1

[15] BK Sriperumbudur, K Fukumizu, A Gretton, GRG Lanckriet, and B Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In Y Bengio, D Schuurmans, J Lafferty, C Williams, and A Culotta, editors, *23rd Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1750–1758, Vancouver, BC, Canada, 2009. Curran. 1

[16] BK Sriperumbudur, K Fukumizu, A Gretton, B Schölkopf, and GRG Lanckriet. Non-parametric estimation of integral probability metrics. In *IEEE International Symposium on Information Theory (ISIT 2010)*, pages 1428–1432, Austin, TX, USA, 6 2010. IEEE. 1

[17] K Zhang, J Peters, D Janzing, and B Schölkopf. Kernel-based conditional independence test and application in causal discovery. In FG Cozman and A Pfeffer, editors, *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813, Barcelona, Spain, 7 2011. AUAI Press. 2