Kernel Mean Embedding of Distributions: A Review and Beyond

Krikamol Muandet
Mahidol University, Thailand
Max Planck Institute for Intelligent Systems, Germany

Kenji Fukumizu
Institute of Statistical Mathematics, Japan

Bharath Sriperumbudur
Pennsylvania State University, USA

Bernhard Schölkopf
Max Planck Institute for Intelligent Systems, Germany

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

K. Muandet, K. Fukumizu, B. Sriperumbudur and B. Schölkopf. *Kernel Mean Embedding of Distributions: A Review and Beyond*. Foundations and Trends® in Machine Learning, vol. 10, no. 1-2, pp. 1–144, 2017.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-288-4

Foundations and Trends® in Machine Learning
Volume 10, Issue 1-2, 2017

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles on the theory, algorithms and applications of machine learning, including the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive, and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2017, Volume 10, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Foundations and Trends® in Machine Learning
Vol. 10, No. 1-2 (2017) 1–141

DOI: 10.1561/2200000060

Kernel Mean Embedding of Distributions: A Review and Beyond

Krikamol Muandet
Department of Mathematics
Faculty of Science, Mahidol University
272 Rama VI Road, Ratchathewi, Bangkok 10400, Thailand

Empirical Inference Department
Max Planck Institute for Intelligent Systems
Spemannstraße 38, Tübingen 72076, Germany
krikamol.mua@mahidol.ac.th

Kenji Fukumizu
Institute of Statistical Mathematics
10-3 Midoricho, Tachikawa, Tokyo 190-8562 Japan
fukumizu@ism.ac.jp

Bharath Sriperumbudur
Department of Statistics, Pennsylvania State University
University Park, PA 16802, USA
bks18@psu.edu

Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
Spemannstraße 38, Tübingen 72076, Germany
bs@tuebingen.mpg.de

Contents

Abstract

A Hilbert space embedding of a distribution—in short, a kernel mean embedding—has recently emerged as a powerful tool for machine learning and statistical inference. The basic idea behind this framework is to map distributions into a reproducing kernel Hilbert space (RKHS) in which the whole arsenal of kernel methods can be extended to probability measures. It can be viewed as a generalization of the original "feature map" common to support vector machines (SVMs) and other kernel methods. In addition to the classical applications of kernel methods, the kernel mean embedding has found novel applications in fields ranging from probabilistic modeling to statistical inference, causal discovery, and deep learning.

This survey aims to give a comprehensive review of existing work and recent advances in this research area, and to discuss challenging issues and open problems that could potentially lead to new research directions. The survey begins with a brief introduction to the RKHS and positive definite kernels which forms the backbone of this survey, followed by a thorough discussion of the Hilbert space embedding of marginal distributions, theoretical guarantees, and a review of its applications. The embedding of distributions enables us to apply RKHS methods to probability measures which prompts a wide range of applications such as kernel two-sample testing, independent testing, and learning on distributional data. Next, we discuss the Hilbert space embedding for conditional distributions, give theoretical insights, and review some applications. The conditional mean embedding enables us to perform sum, product, and Bayes' rules—which are ubiquitous in graphical model, probabilistic inference, and reinforcement learning—in a non-parametric way using this new representation of distributions. We then discuss relationships between this framework and other related areas. Lastly, we give some suggestions on future research directions.

K. Muandet, K. Fukumizu, B. Sriperumbudur and B. Schölkopf. *Kernel Mean Embedding of Distributions: A Review and Beyond*. Foundations and Trends® in Machine Learning, vol. 10, no. 1-2, pp. 1–141, 2017.

DOI: 10.1561/2200000060.

Notation

We summarize a collection of the commonly used notation and symbols in Table 1.

Table 1: Notation and symbols

Symbol	Description
x	A scalar quantity
$\mathbf{x}$	A vector
$\mathbf{X}$	A matrix
X	A random variable
$\mathbb{R}$	Real line or the field of real numbers
$\mathbb{R}^d$	Euclidean d-space
$\mathbb{C}$	Complex plane or the field of complex numbers
$\mathbb{C}^d$	Complex d-space
$\langle \cdot, \cdot \rangle$	An inner product
$\| \cdot \|$	A norm
$\mathcal{X}$, $\mathcal{Y}$	Non-empty spaces in which X and Y take values
$\mathbb{R}^{\mathcal{X}}$	A vector space of functions from $\mathcal{X}$ to $\mathbb{R}$
$\mathscr{H}$	Reproducing kernel Hilbert spaces (RKHS) of functions from $\mathcal{X}$ to $\mathbb{R}$
$\mathscr{G}$	Reproducing kernel Hilbert spaces (RKHS) of functions from $\mathcal{Y}$ to $\mathbb{R}$
$\mathscr{G} \otimes \mathscr{H}$	Tensor product space
$k(\cdot, \cdot)$	Positive definite kernel function on $\mathcal{X} \times \mathcal{X}$
$l(\cdot, \cdot)$	Positive definite kernel function on $\mathcal{Y} \times \mathcal{Y}$
$\phi(\cdot)$	Feature map from $\mathcal{X}$ to $\mathscr{H}$ associated with the kernel k

Table 1: Notation and symbols

Symbol	Description
$\varphi(\cdot)$	Feature map from $\mathcal{Y}$ to $\mathscr{G}$ associated with the kernel l
$\mathbf{K}$	Gram matrix with $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$
$\mathbf{L}$	Gram matrix with $\mathbf{L}_{ij} = l(\mathbf{y}_i, \mathbf{y}_j)$
$\mathbf{H}$	Centering matrix
$\mathbf{K} \circ \mathbf{L}$	Hadamard product of matrices $\mathbf{K}$ and $\mathbf{L}$
$\mathbf{K} \otimes \mathbf{L}$	Kronecker product of matrices $\mathbf{K}$ and $\mathbf{L}$
$\mathcal{C}_{XX}$, $\mathcal{C}_{YY}$	Covariance operators in $\mathscr{H}$ and $\mathscr{G}$, respectively
$\mathcal{C}_{XY}$	Cross-covariance operators from $\mathscr{G}$ to $\mathscr{H}$
$\mathcal{C}_{XY\|Z}$	Conditional cross-covariance operator
$\mathcal{V}_{YX}$	Normalized cross-covariance operator from $\mathscr{H}$ to $\mathscr{G}$
$\mathcal{V}_{YX\|Z}$	Normalized conditional cross-covariance operator
$C_b(\mathcal{X})$	Space of bounded continuous functions on $\mathcal{X}$
$L_2[a,b]$	Space of square-integrable functions on $[a,b]$
$L_2(\mathcal{X},\mu)$	Space of square μ-integrable functions on $\mathcal{X}$
$M^1_+(\mathcal{X})$	Space of probability measures on $\mathcal{X}$
ℓ^1	Space of sequences whose series is absolutely convergent
ℓ^2	Space of square summable sequences
ℓ^∞	Space of bounded sequences
$H^r_2(\mathbb{R}^d)$	Sobolev space of r-times differentiable functions
$\mathrm{HS}(\mathscr{G},\mathscr{H})$	Hilbert space of Hilbert-Schmidt operators mapping from $\mathscr{G}$ to $\mathscr{H}$
$\mathfrak{F}g$	Fourier transform of g
Id	Identity operator
Λ	Spectral density
$\mathcal{N}(\mathcal{C})$	Null space of an operator $\mathcal{C}$
$\mathcal{R}(\mathcal{C})$	Range of an operator $\mathcal{C}$
$S^\perp$	Orthogonal complement of a closed subspace S
$(h_i)_{i\in I}$	Orthonormal basis
$\mathbb{P}$, $\mathbb{Q}$	Probability distributions
$\varphi_{\mathbb{P}}$	Characteristic function of the distribution $\mathbb{P}$
$O(n)$	Order n time complexity of an algorithm
$O_p(n)$	Order n in probability (or stochastic boundedness)

1

Introduction

This work aims to provide a comprehensive review of kernel mean embeddings of distributions and, in the course of doing so, discusses some challenging issues that could potentially lead to new research directions. To the best of our knowledge, there is no comparable review in this area so far; however, the short review paper of Song et al. (2013) on Hilbert space embedding of conditional distributions and its applications in nonparametric inference in graphical models may be of interest to some readers.

The kernel mean embedding owes its success to a positive definite function commonly known as the *kernel function*. The kernel function has become popular in the machine learning community for more than 20 years. Initially, it arises as an effortless way to perform an inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ in a high-dimensional feature space $\mathscr{H}$ for some data points $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. The positive definiteness of the kernel function guarantees the existence of a dot product space $\mathscr{H}$ and a mapping $\phi : \mathcal{X} \to \mathscr{H}$ such that $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathscr{H}}$ (Aronszajn 1950) without needing to compute ϕ explicitly (Boser et al. 1992, Cortes and Vapnik 1995, Vapnik 2000, Schölkopf and Smola 2002). The kernel function can be applied to any learning algorithm as long as the latter can be expressed

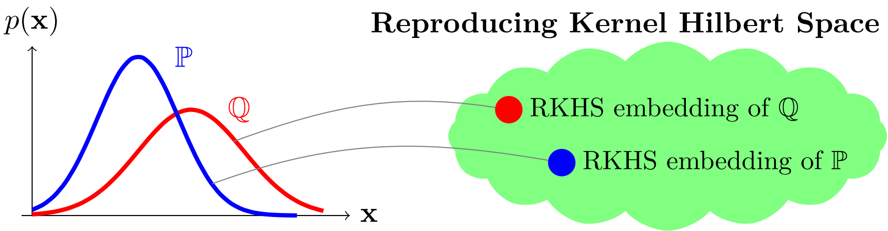

Figure 1.1: Embedding of marginal distributions: each distribution is mapped into a reproducing kernel Hilbert space via an expectation operation.

entirely in terms of a dot product $\langle \mathbf{x}, \mathbf{y} \rangle$ (Schölkopf et al. 1998). This trick is commonly known as the *kernel trick* (see Section 2 for a more detailed account). Many kernel functions have been proposed for various kinds of data structures including non-vectorial data such as graphs, text documents, semi-groups, and probability distributions (Schölkopf and Smola 2002, Gärtner 2003). Many well-known learning algorithms have already been *kernelized* and have proven successful in scientific disciplines such as bioinformatics, natural language processing, computer vision, robotics, and causal inference.

Figures 1.1 and 1.2 depict schematic illustrations of the kernel mean embedding framework. In words, the idea of *kernel mean embedding* is to extend the feature map ϕ to the space of probability distributions by representing each distribution $\mathbb{P}$ as a mean function

$$\phi(\mathbb{P}) = \mu_{\mathbb{P}} := \int_{\mathcal{X}} k(\mathbf{x}, \cdot)\, \mathrm{d}\mathbb{P}(\mathbf{x}), \tag{1.1}$$

where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric and positive definite kernel function (Berlinet and Thomas-Agnan 2004, Smola et al. 2007). Since $k(\mathbf{x}, \cdot)$ takes values in the feature space $\mathscr{H}$, the integral in (1.1) should be interpreted as a Bochner integral (see, *e.g.*, Diestel and Uhl 1977; Chapter 2 and Dinculeanu 2000; Chapter 1 for a definition of the Bochner integral). Conditions ensuring the existence of such an integral will be discussed in Section 3, but in this case we essentially transform the distribution $\mathbb{P}$ to an element in $\mathscr{H}$, which is nothing but a reproducing kernel Hilbert space (RKHS) corresponding to the kernel k. Through (1.1), most RKHS methods can be extended to probability measures. This representation is beneficial for the following reasons.

First of all, for a class of kernel functions known as *characteristic kernels*, the kernel mean representation captures all information about the distribution $\mathbb{P}$ (Fukumizu et al. 2004, Sriperumbudur et al. 2008; 2010). In other words, the mean map $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective, implying that $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathscr{H}} = 0$ if and only if $\mathbb{P} = \mathbb{Q}$, *i.e.*, $\mathbb{P}$ and $\mathbb{Q}$ are the same distribution. Consequently, the kernel mean representation can be used to define a metric over the space of probability distributions (Sriperumbudur et al. 2010). Since $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathscr{H}}$ can be bounded from above by some popular probability metrics such as the Wasserstein distance and the total variation distance, it follows that if $\mathbb{P}$ and $\mathbb{Q}$ are close in these distances, then $\mu_{\mathbb{P}}$ is also close to $\mu_{\mathbb{Q}}$ in the $\|\cdot\|_{\mathscr{H}}$ norm (see §3.5). Injectivity of $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ makes it suitable for applications that require a unique characterization of distributions such as two-sample homogeneity tests (Gretton et al. 2012a, Fukumizu et al. 2008, Zhang et al. 2011, Doran et al. 2014). Moreover, using the kernel mean representation, most learning algorithms can be extended to the space of probability distributions with minimal assumptions on the underlying data generating process (Gómez-Chova et al. 2010, Muandet et al. 2012, Guevara et al. 2015, Lopez-Paz et al. 2015). See §3.3 for details.

Secondly, several elementary operations on distributions (and associated random variables) can be performed directly by means of this representation. For example, by the reproducing property of $\mathscr{H}$,

$$\mathbb{E}_{\mathbb{P}}[f(\mathbf{x})] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathscr{H}}, \quad \forall f \in \mathscr{H}.$$

That is, an expected value of any function $f \in \mathscr{H}$ w.r.t. $\mathbb{P}$ is nothing but an inner product in $\mathscr{H}$ between f and $\mu_{\mathbb{P}}$. Likewise, for an RKHS $\mathscr{G}$ over some input space $\mathcal{Y}$, we have

$$\mathbb{E}_{Y|\mathbf{x}}[g(Y) \mid X = \mathbf{x}] = \langle g, \mathcal{U}_{Y|\mathbf{x}} \rangle_{\mathscr{G}}, \quad \forall g \in \mathscr{G},$$

where $\mathcal{U}_{Y|\mathbf{x}}$ denotes the embedding of the conditional distribution $\mathbb{P}(Y|X = \mathbf{x})$. That is, we can compute a conditional expected value of any function $g \in \mathscr{G}$ w.r.t. $\mathbb{P}(Y|X = \mathbf{x})$ by taking an inner product in $\mathscr{G}$ between the function g and the embedding of $\mathbb{P}(Y|X = \mathbf{x})$ (see Section 4 for further details). These operations only require knowledge of the empirical estimates of $\mu_{\mathbb{P}}$ and $\mathcal{U}_{Y|\mathbf{x}}$. Hence, the kernel mean representation allows us to implement these operations in *non-parametric* probabilistic inference, *e.g.*, filtering for dynamical systems (Song et al. 2009), kernel belief propagation (Song et al. 2011a), kernel Monte Carlo filter (Kanagawa et al. 2013), kernel Bayes' rule (Fukumizu et al. 2011), often with strong theoretical guarantees. Moreover, it can be used to perform functional operations $f(X, Y)$ on random variables X and Y (Schölkopf et al. 2015, Simon-Gabriel et al. 2016).

In some applications such as testing for homogeneity from finite samples, representing the distribution $\mathbb{P}$ by $\mu_{\mathbb{P}}$ bypasses an intermediate density estimation, which is known to be difficult in the high-dimensional setting (Wasserman 2006; Section 6.5). Moreover, we can extend the applications of kernel mean embedding straightforwardly to non-vectorial data such as graphs, strings, and semi-groups, thanks to the kernel function. As a result, statistical inference—such as two-sample testing and independence testing—can be adapted directly to distributions over complex objects (Gretton et al. 2012a).

Under additional assumptions, we can generalize the principle underlying (1.1) to conditional distributions $\mathbb{P}(Y|X)$ and $\mathbb{P}(Y|X = \mathbf{x})$. Essentially, the latter two objects are represented as an operator that maps the feature space $\mathscr{H}$ to $\mathscr{G}$, and as an object in the feature space $\mathscr{G}$, respectively, where $\mathscr{H}$ and $\mathscr{G}$ denote the RKHS for X and Y, respectively (see Figure 1.2). These representations allow us to develop a powerful language for algebraic manipulation of probability distributions in an analogous way to the sum rule, product rule, and Bayes' rule—which are ubiquitous in graphical models and probabilistic inference—without making assumption on parametric forms of the underlying distributions. The details of conditional mean embeddings will be given in Section 4.

A Synopsis. As a result of the aforementioned advantages, the kernel mean embedding has made widespread contributions in various directions. Firstly, most tasks in machine learning and statistics involve estimation of the data-generating process whose success depends critically on the accuracy and the reliability of this estimation. It is known that estimating the kernel mean embedding is easier than estimating the distribution itself, which helps improve many statistical inference

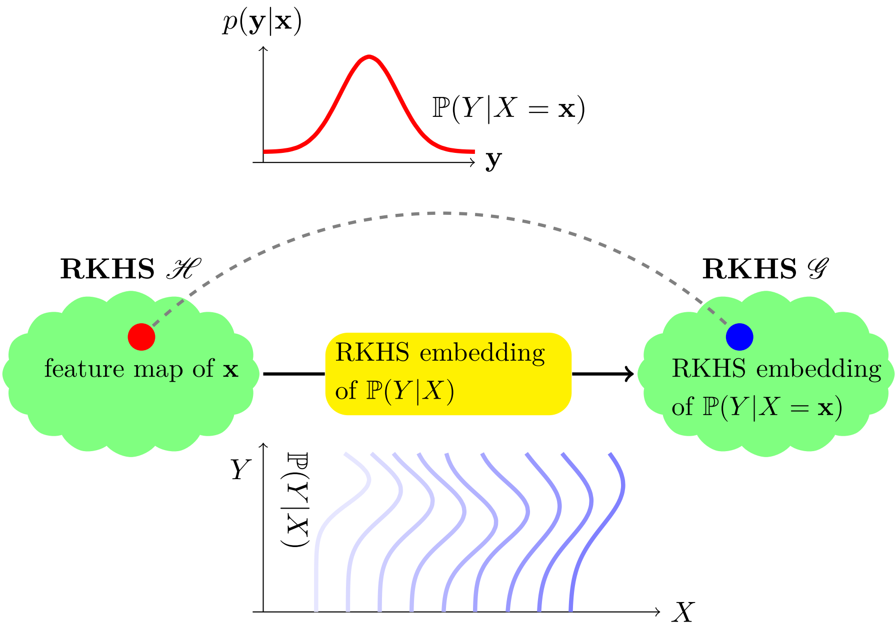

Figure 1.2: From marginal distribution to conditional distribution: unlike the embeddings shown in Figure 1.1, the embedding of conditional distribution $\mathbb{P}(Y|X)$ is not a single element in the RKHS. Instead, it may be viewed as a family of Hilbert space embeddings of conditional distributions $\mathbb{P}(Y|X=\mathbf{x})$ indexed by the conditioning variable X. In other words, the conditional mean embedding can be viewed as an operator from $\mathscr{H}$ to $\mathscr{G}$ (cf. §4.2).

methods. These include, for example, two-sample testing (Gretton et al. 2012a), independence and conditional independence tests (Fukumizu et al. 2008, Zhang et al. 2011, Doran et al. 2014), causal inference (Sgouritsa et al. 2013, Chen et al. 2014), adaptive MCMC (Sejdinovic et al. 2014), and approximate Bayesian computation (Fukumizu et al. 2013).

Secondly, several attempts have been made in using kernel mean embedding as a representation in the predictive learning on distributions (Muandet et al. 2012, Szabó et al. 2016, Muandet and Schölkopf 2013, Guevara et al. 2015, Lopez-Paz et al. 2015). As opposed to the classical setting where training and test examples are data points, many applications call for a learning framework in which training and test examples are probability distributions. This is ubiquitous in, for example, multiple-instance learning (Doran 2013), learning with noisy and uncertain input, learning from missing data, group anomaly detection (Muandet and Schölkopf 2013, Guevara et al. 2015), dataset squishing, and bag-of-words data (Yoshikawa et al. 2014; 2015). The kernel mean representation equipped with the RKHS methods enables classification, regression, and anomaly detection to be performed on such distributions.

Finally, the kernel mean embedding also allows one to perform complex approximate inference without making strong parametric assumption on the form of underlying distribution. The idea is to represent all relevant probabilistic quantities as a kernel mean embedding. Then, basic operations such as *sum rule* and *product rule* can be formulated in terms of the expectation and inner product in feature space. Examples of algorithms in this class include kernel belief propagation (KBP), kernel embedding of latent tree model, kernel Bayes rule, and predictive-state representation (Song et al. 2010b; 2009; 2011a; 2013, Fukumizu et al. 2013). Recently, the kernel mean representation has become one of the prominent tools in causal inference and discovery (Lopez-Paz et al. 2015, Sgouritsa et al. 2013, Chen et al. 2014, Schölkopf et al. 2015).

The aforementioned examples represent only a handful of successful applications of kernel mean embedding. More examples and details will be provided throughout the survey.

1.1 Purpose and Scope

The purpose of this survey is to give a comprehensive review of kernel mean embedding of distributions, to present important theoretical results and practical applications, and to draw connections to related areas. We restrict the scope of this survey to key theoretical results and new applications of kernel mean embedding with references to related work. We focus primarily on basic intuition and sketches for proofs, leaving the full proofs to the papers cited.

All materials presented in this paper should be accessible to a wide audience. In particular, we hope that this survey will be most useful to readers who are not at all familiar with the idea of kernel mean embedding, but already have some background knowledge in machine learning. To ease the reading, we suggest non-expert readers to also consult elementary machine learning textbooks such as Bishop (2006), Schölkopf and Smola (2002), Mohri et al. (2012), and Murphy (2012). Experienced machine learners who are interested in applying the idea of kernel mean embedding to their work are also encouraged to read this survey. Lastly, we will also provide some practical considerations that could be useful to practitioners who are interested in implementing the idea in real-world applications.

1.2 Outline of the Survey

The schematic outline of this survey is depicted in Figure 1.3 and can be summarized as follows.

Section 2 introduces notation and the basic idea of a positive definite kernel and reproducing kernel Hilbert space (RKHS) (§2.1 and §2.2). It also presents general theoretical results such as the reproducing property (Prop 2.1), the Riesz representation theorem (Thm 2.4), Mercer's theorem (Thm 2.1), Bochner's theorem (Thm 2.2), and Schoenberg's characterization (Thm 2.3). In addition, it contains a brief discussion about Hilbert-Schmidt operators on RKHS (§2.3).

Section 3 conveys the idea of Hilbert space embedding of marginal distributions (§3.1) as well as covariance operators (§3.2), presents essential properties of mean embedding (§3.3), discusses its estimation and approximation procedures (§3.4), and reviews important applications, notably maximum mean discrepancy (MMD) (§3.5), kernel dependence measure (§3.6), learning on distributional data (§3.7), and how to recover information from the embedding of distributions (§3.8).

Section 4 generalizes the idea of kernel mean embedding to the space of conditional distributions, called *conditional mean embedding* (§4.1), presents regression perspective (§4.2), and describes basic operations—namely sum rule, product rule, and Bayes' rule—in terms of marginal

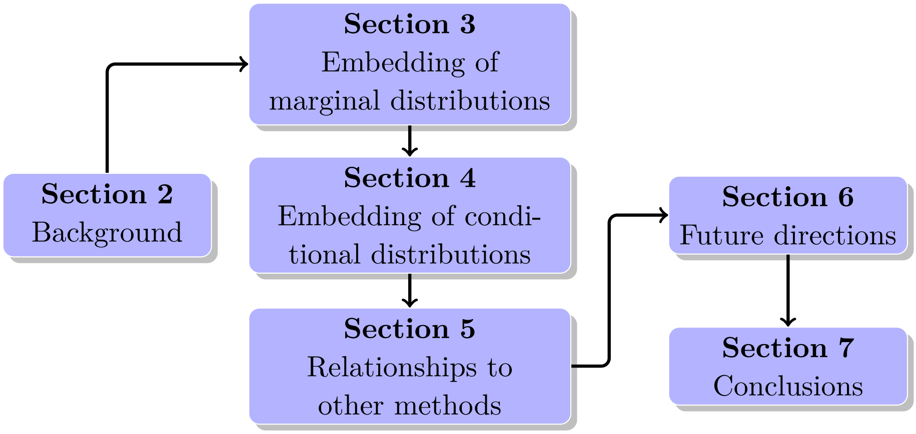

Figure 1.3: Schematic outline of this survey.

and conditional mean embeddings (§4.3). We review applications in graphical models, probabilistic inference (§4.4), reinforcement learning (§4.6), conditional dependence measures (§4.7), and causal discovery (§4.8). Estimating the conditional mean embedding is challenging both theoretically and empirically. We discuss some of the key challenges as well as some applications.

Section 5 draws connections between the kernel mean embedding framework and other methods including kernel density estimation, empirical characteristic function, divergence methods and probabilistic modeling. Section 6 provides suggestions for future research. Lastly, Section 7 concludes the survey.

References

S. Achard, D. Pham, and C. Jutten. Quadratic dependence measure for nonlinear blind source separation. In *Proceeding of 4th International Symposium on Independent Component Analysis and Blind Source Separation (ICA2003)*, pages 263–268, 2003.

R. Adams. *Kernel Methods for Nonparametric Bayesian Inference of Probability Densities and Point Processes*. PhD thesis, University of Cambridge, 2009.

R. Adams and J. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics*. Elsevier/Academic Press, Amsterdam, 2nd edition, 2003.

N. Akhiezer and I. Glazman. *Theory of linear operators in Hilbert space*. Dover Publications Inc., New York, 1993.

V. Alba Fernàndez, M. Jimènez Gamero, and J. Muñoz Garcia. A test for the two-sample problem based on empirical characteristic functions. *Computational Statistics & Data Analysis*, 52(7):3730–3748, 2008.

M. Álvarez, L. Rosasco, and N. Lawrence. Kernels for vector-valued functions: A review. *Foundation and Trends in Machine Learning*, 4(3):195–266, 2012.

H. Anderson and M. Gupta. Expected kernel for missing features in support vector machines. In *Statistical Signal Processing Workshop*, pages 285–288, 2011.

N. Anderson, P. Hall, and D. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50 (1):41–54, July 1994.

C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373, 2008.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

K. Baba, R. Shibata, and M. Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.

F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *The 26th Annual Conference on Learning Theory*, volume 30 of *JMLR Proceedings*, pages 185–209. JMLR.org, 2013.

F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. Technical report, INRIA, 2015.

F. Bach and M. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2003.

F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1359–1366, 2012.

C. Baker. Mutual information for Gaussian processes. *SIAM Journal on Applied Mathematics*, 19(2):451–458, 1970.

C. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

L. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

F. Bavaud. On the Schoenberg transformations in data analysis: Theory and illustrations. *Journal of Classification*, 28(3):297–314, 2011.

R. Bellman. A Markovian decision process. *Indiana University Mathematics Journal*, 6:679–684, 1957.

R. Bellman. *Dynamic Programming*. Dover Publications, Incorporated, 2003.

J. Berger and R. Wolpert. Estimating the mean function of a Gaussian process and the Stein effect. *Journal of Multivariate Analysis*, 13(3):401–424, 1983.

W. Bergsma. Testing conditional independence for continuous random variables. *EURANDOM-report*, 2004.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

M. Besserve, N. Logothetis, and B. Schölkopf. Statistical analysis of coupled time series with kernel cross-spectral density operators. In *Advances in Neural Information Processing Systems 26*, pages 2535–2543. Curran Associates, Inc., 2013.

A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of Calcutta Mathematical Society*, 35(1):99–109, 1943.

G. Biau and L. Györfi. On the asymptotic properties of a nonparametric l/sub 1/-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11):3965–3973, 2005.

C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems*, pages 2178–2186. 2011.

G. Blom. Some properties of incomplete U-statistics. *Biometrika*, 63(3):573–580, 1976.

A. Blum. Random projection, margins, kernels, and feature-selection. In *Proceedings of the International Conference on Subspace, Latent Structure and Feature Selection*, volume 3940 of *Lecture Notes in Computer Science*, pages 52–68. Springer-Verlag, 2005.

S. Bochner. Monotone funktionen, Stieltjessche integrale und harmonische analyse. *Mathematische Annalen*, 108(1):378–410, 1933.

B. Boots, A. Gretton, and G. Gordon. Hilbert space embeddings of predictive state representations. In *Proceedings of the 29th International Conference on Uncertainty in Artificial Intelligence*, pages 92–101, 2013.

K. Borgwardt, A. Gretton, M. Rasch, H.-P. Kriegel, B. Schölkopf, and A. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):49–57, 2006.

B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 144–152. ACM, 1992.

D. Bouchacourt, P. Mudigonda, and S. Nowozin. DISCO Nets : DISsimilarity COefficients networks. In *Advances in Neural Information Processing Systems 29*, pages 352–360. Curran Associates, Inc., 2016.

W. Bounliphone, A. Gretton, A. Tenenhaus, and M. Blaschko. A low variance consistent test of relative dependency. In *Proceedings of the 32nd International Conference on Machine Learning*, JMLR Proceedings, pages 20–29. JMLR.org, 2015.

W. Bounliphone, E. Belilovsky, M. Blaschko, I. Antonoglou, and A. Gretton. A test of relative similarity for model selection in generative models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, May 2016.

C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, June 1998.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

A. Caponnetto, C. Micchelli, M. Pontil, and Y. Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9:1615–1646, 2008.

C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 04(04):377–408, 2006.

O. Chapelle, J. Weston, L. Bottou, and V. Vapnik. Vicinal risk minimization. In *Advances in Neural Information Processing Systems 13*, pages 416–422. MIT Press, 2001.

Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 109–116. AUAI Press, 2010.

Z. Chen, K. Zhang, L. Chan, and B. Schölkopf. Causal discovery via reproducing kernel Hilbert space embeddings. *Neural Computation*, 26(7): 1484–1517, 2014.

A. Christmann and I. Steinwart. Universal kernels on non-standard input spaces. In *Advances in Neural Information Processing Systems*, pages 406–414. 2010.

K. Chwialkowski and A. Gretton. A kernel independence test for random processes. In *Proceedings of The 31st International Conference on Machine Learning*, volume 32, pages 1422–1430, 2014.

K. Chwialkowski, D. Sejdinovic, and A. Gretton. A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems 27*, pages 3608–3616, 2014.

K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representation of probability measures. In *Advances in Neural Information Processing Systems 28*, pages 1972–1980, 2015.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *Proceedings of the 33nd International Conference on Machine Learning*, pages 2606–2615, 2016.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, 1995.

C. Cortes, M. Mohri, and J. Weston. A general regression technique for learning transductions. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 153–160. ACM, 2005.

E. Cruz Cortés and C. Scott. Scalable sparse approximation of a sample mean. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pages 5274–5278, 2014.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

F. Cucker and D.-X. Zhou. *Learning Theory: An approximation theory viewpoint.* Cambridge Monographs on Applied and Computational Mathematic. Cambridge University Press, 2007.

M. Cuturi, K. Fukumizu, and J.-P. Vert. Semigroup kernels on measures. *Journal of Machine Learning Research*, 6:1169–1198, 2005.

S. Danafar, P. Rancoita, T. Glasmachers, K. Whittingstall, and J. Schmidhuber. Testing hypotheses by regularized maximum mean discrepancy. *International Journal of Computer and Information Technology*, 3:223–232, 2014.

S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, Jan. 2003.

J. Daudin. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590, 1980.

B. Davies. *Linear Operators and Their Spectra.* Cambridge University Press, 2007.

A. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979.

P. Dayan and L. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.* The MIT Press, 2005.

E. De Vito, L. Rosasco, and R. Verri. Spectral methods for regularization in learning theory. Technical Report DISI-TR-05-18, Universitá di Genova, 2006.

I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 551–556, 2004.

J. Diestel and J. Uhl. *Vector Measures*. American Mathematical Society, Providence, 1977.

N. Dinculeanu. *Vector Integration and Stochastic Integration in Banach Spaces*. Wiley, 2000.

G. Doran. Distribution kernel methods for multiple-instance learning. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 1660–1661. AAAI Press, 2013.

G. Doran, K. Muandet, K. Zhang, and B. Schölkopf. A permutation-based kernel conditional independence test. In *30th Conference on Uncertainty in Artificial Intelligence*, pages 132–141, 2014.

R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. A Wiley Interscience Publication. Wiley, 1973.

C. Dwork. Differential privacy: A survey of results. In *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation (TAMC2008)*, pages 1–19. Springer-Verlag, 2008.

G. Dziugaite, R. Daniel, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 258–267, 2015.

Y. Engel, S. Mannor, and R. Meir. Bayes meets Bellman: The Gaussian process approach to temporal difference learning. In *Proceedings of the 20th International Conference on Machine Learning*, pages 154–161. AAAI Press, 2003.

H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Mathematics and Its Applications (Book 375). Springer, 1996.

S. M. A. Eslami, D. Tarlow, P. Kohli, and J. Winn. Just-in-time learning for fast and flexible inference. In *Advances in Neural Information Processing Systems 27*, pages 154–162. Curran Associates, Inc., 2014.

A. Feuerverger and R. Mureika. The empirical characteristic function and its applications. *The Annals of Statistics*, 5(1):88–97, 1977.

S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

S. Flaxman, D. Neill, and A. Smola. Gaussian processes for independence tests with non-iid data in causal inference. *ACM Transactions on Intelligent Systems and Technology*, 7(2):22:1–22:23, 2015a.

S. Flaxman, Y.-X. Wang, and A. Smola. Who supported Obama in 2012?: Ecological inference through distribution regression. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 289–298. ACM, 2015b.

S. Flaxman, D. Sejdinovic, J. Cunningham, and S. Filippi. Bayesian learning of kernel embeddings. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, pages 182–191. AUAI Press, 2016.

G. Folland. *Real analysis.* Pure and Applied Mathematics. John Wiley & Sons, Inc., New York, 2nd edition, 1999.

K. Fukumizu. *Modern Methodology and Applications in Spatial-Temporal Modeling*, chapter Nonparametric Bayesian Inference with Kernel Mean Embedding, pages 1–24. Springer Japan, Tokyo, 2015.

K. Fukumizu, F. Bach, and M. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

K. Fukumizu, F. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8: 361–383, 2007.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496. Curran Associates, Inc., 2008.

K. Fukumizu, F. Bach, and M. Jordan. Kernel dimension reduction in regression. *Annals of Statistics*, 37(4):1871–1905, 2009a.

K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf. Characteristic kernels on groups and semigroups. In *Advances in Neural Information Processing Systems 21*, pages 473–480. Curran Associates, Inc., 2009b.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule. In *Advances in Neural Information Processing Systems*, pages 1737–1745. 2011.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14: 3753–3783, 2013.

T. Gärtner. A survey of kernels for structured data. *SIGKDD Explorations Newsletter*, 5(1):49–58, July 2003.

T. Gärtner, P. Flach, A. Kowalczyk, and A. Smola. Multi-instance kernels. In *In Proceeding of the 19th International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 2002.

M. Genton. Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, 2:299–312, 2002.

A. Girard, C. Rasmussen, J. Candela, and R. Murray-Smith. Gaussian process priors with uncertain inputs - application to multiple-step ahead time series forecasting. In *Advances in Neural Information Processing Systems 15*, pages 529–536, 2002.

L. Gómez-Chova, G. Camps-Valls, L. Bruzzone, and J. Calpe-Maravilla. Mean map kernel methods for semisupervised cloud classification. *IEEE Transaction on Geoscience and Remote Sensing*, 48(1-1):207–220, 2010.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

N. Goodman, T. Ullman, and J. Tenenbaum. Learning a theory of causality. *Psychological Review*, 2011.

A. Gordon, T. Henzinger, A. Nori, and S. Rajamani. Probabilistic programming. In *Proceedings of the on Future of Software Engineering*, FOSE 2014, pages 167–181, New York, NY, USA, 2014. ACM.

A. Gretton. Reproducing kernel Hilbert spaces in machine learning. http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/rkhscourse.html, 2016.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-schmidt norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, pages 63–77. Springer-Verlag, 2005a.

A. Gretton, R. Herbrich, A. Smola, B. Schölkopf, and A. Hyvärinen. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005b.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520. MIT Press, 2007.

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. *Covariate Shift by Kernel Mean Matching*, pages 131–160. MIT Press, 2009.

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

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 *Advances in Neural Information Processing Systems*, pages 1214–1222, 2012b.

S. Grünewälder, G. Lever, L. Baldassarre, M. Pontil, and A. Gretton. Modelling transition dynamics in MDPs with RKHS embeddings. In *Proceedings of the 29th International Conference on Machine Learning*, pages 535–542. Omnipress, 2012.

S. Grünewälder, G. Lever, A. Gretton, L. Baldassarre, S. Patterson, and M. Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1823–1830. Omnipress, 2012.

S. Grünewälder, A. Gretton, and J. Shawe-Taylor. Smooth operators. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1184–1192, 2013.

J. Guevara, S. Canu, and R. Hirata. Support Measure Data Description for group anomaly detection. In *ODDx3 Workshop on Outlier Definition, Detection, and Description at the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, Australia, 2015.

I. Guyon. Cause-effect pairs Kaggle competition, 2013. URL https://www.kaggle.com/c/cause-effect-pairs/.

I. Guyon. ChaLearn fast causation coefficient challenge, 2014. URL https://www.codalab.org/competitions/1381.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, Mar. 2003.

H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

J. Hammersley and P. Clifford. Markov random fields on finite graphs and lattices. Unpublished manuscript, 1971.

Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In *Advances in Neural Information Processing Systems 20*, pages 609–616. Curran Associates, Inc., 2007.

Z. Harchaoui, E. Moulines, and F. Bach. Kernel change-point analysis. In *Advances in Neural Information Processing Systems 21*, pages 609–616. Curran Associates, Inc., 2009a.

Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, and O. Cappe. A regularized kernel-based approach to unsupervised audio segmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1665–1668, 2009b.

S. Harmeling, M. Hirsch, and B. Schölkopf. On a link between kernel mean maps and Fraunhofer diffraction, with an application to super-resolution beyond the diffraction limit. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1083–1090. IEEE, 2013.

D. Haussler. Convolution kernels on discrete structures. Technical Report UCS-CRL-99-10, University of California at Santa Cruz, Santa Cruz, CA, USA, 1999.

M. Hein and O. Bousquet. Kernels, associated structures and generalizations. Technical Report 127, Max-Planck-Gesellschaft, July 2004.

M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pages 136–143, 2005.

J. Hodges and E. Lehmann. The efficiency of some nonparametric competitors of the t-test. *The Annals of Mathematical Statistics*, 27(2):324–335, 1956.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

T. Hofmann, B. Schölkopf, and A. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

D. Hsu, S. Kakade, and T. Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, pages 601–608. MIT Press, 2007.

P. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

F. Huszar and D. Duvenaud. Optimally-weighted herding is Bayesian quadrature. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 377–386. AUAI Press, 2012.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

A. Ihler and D. McAllester. Particle belief propagation. *International Conference on Artificial Intelligence and Statistics*, 5:256–263, 2009.

T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, pages 487–493. MIT Press, 1998.

H. Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000.

W. James and J. Stein. Estimation with quadratic loss. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379. University of California Press, 1961.

D. Janzing, E. Sgouritsa, O. Stegle, J. Peters, and B. Schölkopf. Detecting low-complexity unobserved causes. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 383–391. AUAI Press, 2011.

D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Supplement to: Quantifying causal influences. *The Annals of Statistics*, 2013.

T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

S. Jegelka, A. Gretton, B. Schölkopf, B. Sriperumbudur, and U. von Luxburg. Generalized clustering via kernel embeddings. In *KI 2009: AI and Automation, Lecture Notes in Computer Science, Vol. 5803*, pages 144–152. Springer, 2009.

W. Jitkrittum, A. Gretton, N. Heess, A. Eslami, B. Lakshminarayanan, D. Sejdinovic, and Z. Szabó. Kernel-based just-in-time learning for passing expectation propagation messages. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 405–414. AUAI Press, 2015.

W. Jitkrittum, Z. Szabó, K. Chwialkowski, and A. Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems 29*, pages 181–189. Curran Associates, Inc., 2016.

M. Kanagawa and K. Fukumizu. Recovering distributions from Gaussian RKHS embeddings. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pages 457–465. JMLR, 2014.

M. Kanagawa, Y. Nishiyama, A. Gretton, and K. Fukumizu. Kernel Monte Carlo filter. Master's thesis, 2013.

M. Kanagawa, Y. Nishiyama, A. Gretton, and K. Fukumizu. Filtering with state-observation examples via kernel Monte Carlo filter. *Neural Computation*, 28(2):382–444, 2016a.

M. Kanagawa, B. Sriperumbudur, and K. Fukumizu. Convergence guarantees for kernel-based quadrature rules in misspecified settings. In *Advances in Neural Information Processing Systems 29*, pages 3288–3296. Curran Associates, Inc., 2016b.

A. Kankainen. *Consistent Testing of Total Independence Based on the Empirical Characteristic Function.* PhD thesis, University of Jyvaskyla, 1995.

A. Kankainen and N. Ushakov. A consistent modification of a test for independence based on the empirical characteristic function. *Journal of Mathematical Sciences*, 89(5):1486–1494, 1998.

P. Kar and H. Karnick. Random feature maps for dot product kernels. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22, pages 583–591. JMLR.org, 2012.

B. Kim, O. Koyejo, and R. Khanna. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems 29*, pages 2280–2288. Curran Associates, Inc., 2016.

J. Kim and C. Scott. Robust kernel density estimation. *Journal of Machine Learning Research*, 13:2529–2565, Sep. 2012.

K. Kim, M. Franz, and B. Schölkopf. Iterative kernel principal component analysis for image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1351–1366, 2005.

R. Kondor and T. Jebara. A kernel between sets of vectors. In *Proceedings of the 20th International Conference on Machine Learning*, pages 361–368. AAAI Press, 2003.

S. Kpotufe. k-nn regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems 24*, pages 729–737. Curran Associates, Inc., 2011.

E. Kreyszig. *Introductory Functional Analysis with Application.* Wiley, 1978.

A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2-3):123–286, 2012.

J. Kwok and I. Tsang. The pre-image problem in kernel methods. *IEEE Transactions on Neural Networks*, 15(6):1517–1525, Nov. 2004.

S. Lacoste-Julien, F. Lindsten, and F. Bach. Sequential kernel herding: Frank-Wolfe optimization for particle filtering. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38, pages 544–552, May 2015.

Q. Le, T. Sarlos, and A. Smola. Fastfood–approximating kernel expansions in loglinear time. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 244–252, 2013.

Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1718–1727, 2015.

Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33nd International Conference on Machine Learning*, pages 276–284, 2016.

J. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems 28*, pages 829–837. Curran Associates, Inc., 2015.

M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 97–105, 2015.

M. Long, H. Zhu, J. Wang, and M. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems 29*, pages 136–144. Curran Associates, Inc., 2016.

D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. Towards a learning theory of cause-effect inference. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1452–1461, 2015.

A. Mandelbaum and L. Shepp. Admissibility as a touchstone. *Annals of Statistics*, 15(1):252–268, 1987.

A. Martins, N. Smith, E. Xing, P. Aguiar, and M. Figueiredo. Nonextensive information theoretic kernels on measures. *Journal of Machine Learning Research*, 10:935–975, 2009.

L. McCalman, S. O'Callaghan, and F. Ramos. Multi-modal estimation with kernel embeddings for learning motion models. In *IEEE International Conference on Robotics and Automation*, pages 2845–2852, May 2013.

R. Megginson. *An Introduction to Banach Space Theory.* Springer-Verlag New York, Inc., 1998.

N. Mehta and A. Gray. Generative and latent mean map kernels. *CoRR*, abs/1005.0188, 2010.

J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 209 (441-458):415–446, 1909.

C. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

T. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning.* The MIT Press, 2012.

J. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

P. Moreno, P. Ho, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In *Advances in Neural Information Processing Systems 16*, pages 1385–1392. MIT Press, 2004.

K. Muandet. *From Points to Probability Measures: Statistical Learning on Distributions with Kernel Mean Embedding.* PhD thesis, Department of Computer Science, Tübingen University, 2015.

K. Muandet and B. Schölkopf. One-class support measure machines for group anomaly detection. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pages 449–458. AUAI Press, 2013.

K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems*, pages 10–18. 2012.

K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning*, pages 10–18. JMLR, 2013.

K. Muandet, K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf. Kernel mean estimation and Stein effect. In *Proceedings of The 31st International Conference on Machine Learning*, volume 32, pages 10–18. JMLR, 2014a.

K. Muandet, B. Sriperumbudur, and B. Schölkopf. Kernel mean estimation via spectral filtering. In *Advances in Neural Information Processing Systems 27*, pages 10–18. Curran Associates, Inc., 2014b.

K. Muandet, B. Sriperumbudur, K. Fukumizu, A. Gretton, and B. Schölkopf. Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 17(48):1–41, 2016.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

K. Murphy. *Machine Learning: A Probabilistic Perspective.* The MIT Press, 2012.

R. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, September 1993.

X. Nguyen, M. Wainwright, and M. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems 20*, pages 1089–1096. Curran Associates, Inc., 2007.

Y. Nishiyama and K. Fukumizu. Characteristic kernels and infinitely divisible distributions. *Journal of Machine Learning Research*, 17(180):1–28, 2016.

Y. Nishiyama, A. Boularias, A. Gretton, and K. Fukumizu. Hilbert space embeddings of POMDPs. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 644–653, 2012.

C. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

J. Oliva, B. Póczos, and J. Schneider. Distribution to distribution regression. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *JMLR Proceedings*, pages 1049–1057, 2013.

J. Oliva, B. Poczos, and J. Schneider. Fast distribution to real regression. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, volume 33 of *JMLR Proceedings*, pages 706–714, 2014.

S. Pan, I. Tsang, J. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

M. Park, W. Jitkrittum, and D. Sejdinovic. K2-ABC: Approximate Bayesian computation with infinite dimensional summary statistics via kernel embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 398–407, 2016.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers Inc., 1988.

J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 2000.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

J. Pennington, F. Yu, and S. Kumar. Spherical random features for polynomial kernels. In *Advances in Neural Information Processing Systems 28*, pages 1846–1854. Curran Associates, Inc., 2015.

N. Pham and R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 239–247. ACM, 2013.

B. Póczos, L. Xiong, and J. Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 599–608, 2011.

B. Póczos, L. Xiong, D. Sutherland, and J. Schneider. Nonparametric kernel estimators for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2989–2996, 2012.

B. Póczos, A. Singh, A. Rinaldo, and L. Wasserman. Distribution-free distribution regression. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, volume 31 of *JMLR Proceedings*, pages 507–515, 2013.

J. Porta, N. Vlassis, M. Spaan, and P. Poupart. Point-based value iteration for continuous POMDPs. *Journal of Machine Learning Research*, 7:2329–2367, 2006.

N. Privault and A. Réveillac. Stein estimation for the drift of Gaussian processes using the Malliavin calculus. *Annals of Statistics*, 36(5):2531–2550, 2008.

N. Quadrianto, L. Song, and A. Smola. Kernelized sorting. In *Advances in Neural Information Processing Systems 21*, pages 1289–1296. Curran Associates, Inc., 2009.

M. Quang, M. Biagio, and V. Murino. Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces. In *Advances in Neural Information Processing Systems 27*, pages 388–396. Curran Associates, Inc., 2014.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2007.

A. Ramdas and L. Wehbe. Nonparametric independence testing for small sample sizes. In *Proceedings of the 2015 International Joint Conference on Artificial Intelligence*, pages 3777–3783, 2015.

A. Ramdas, S. Reddi, B. Póczos, A. Singh, and L. Wasserman. Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing. *CoRR*, abs/1508.00655, 2015a.

A. Ramdas, S. Reddi, B. Póczos, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 3571–3577. AAAI Press, 2015b.

C. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems 15*, pages 489–496. MIT Press, 2002.

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning).* The MIT Press, 2005.

S. Reddi, A. Ramdas, B. Poczos, A. Singh, and L. Wasserman. On the high dimensional power of a linear-time two sample test under mean-shift alternatives. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 772–780, 2015.

M. Reed and B. Simon. *Functional Analysis, Volume 1 (Methods of Modern Mathematical Physics).* Academic Press, 1st edition, 1981.

A. Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3-4):441–451, 1959.

C. Robert and G. Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

L. Rosasco, M. Belkin, and E. D. Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11:905–934, 2010.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

W. Rudin. *Functional Analysis.* International Series in Pure and Applied Mathematics. McGraw-Hill Inc., New York, 2nd edition, 1991.

I. Schoenberg. Metric spaces and completely monotone functions. *The Annals of Mathematics*, 39(4):811–841, 1938.

B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2002.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, July 1998.

B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, COLT '01/EuroCOLT '01, pages 416–426. Springer-Verlag, 2001.

B. Schölkopf, K. Tsuda, and J.-P. Vert, editors. *Kernel Methods in Computational Biology.* Computational Molecular Biology. MIT Press, 2004.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1255–1262. Omnipress, 2012.

B. Schölkopf, K. Muandet, K. Fukumizu, and J. Peters. Computing functions of random variables via reproducing kernel Hilbert space representations. *Statistics and Computing*, 25(4):755–766, 2015.

I. Schuster, H. Strathmann, B. Paige, and D. Sejdinovic. Kernel sequential Monte Carlo. *CoRR*, abs/:1510.03105, 2016.

D. Sejdinovic, A. Gretton, B. Sriperumbudur, and K. Fukumizu. Hypothesis testing using pairwise distances and associated kernels. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1111–1118. Omnipress, 2012.

D. Sejdinovic, A. Gretton, and W. Bergsma. A kernel test for three-variable interactions. In *Advances in Neural Information Processing Systems 26*, pages 1124–1132. Curran Associates, Inc., 2013a.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013b.

D. Sejdinovic, H. Strathmann, M. Garcia, C. Andrieu, and A. Gretton. Kernel adaptive Metropolis-Hastings. In *Proceedings of the 31th International Conference on Machine Learning*, pages 1665–1673, 2014.

R. Serfling. *Approximation theorems of mathematical statistics.* Wiley Series in Probability and Statistics. Wiley, 1981.

E. Sgouritsa, D. Janzing, J. Peters, and B. Schölkopf. Identifying finite mixtures of nonparametric product distributions and causal inference of confounders. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pages 556–565. AUAI Press Corvallis, 2013.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

J. Shawe-Taylor and A. Dolia. A framework for probability density estimation. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, volume 2 of *JMLR Proceedings*, pages 468–475, 2007.

P. Shivaswamy, C. Bhattacharyya, and A. Smola. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7:1283–1314, 2006.

B. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London, 1986.

C.-J. Simon-Gabriel, A. Scibior, I. Tolstikhin, and B. Schölkopf. Consistent kernel mean estimation for functions of random variables. In *Advances in Neural Information Processing Systems 29*, pages 1732–1740. Curran Associates, Inc., 2016.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 2007.

R. Smallwood and E. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer-Verlag, 2007.

L. Song. *Learning via Hilbert Space Embedding of Distributions.* PhD thesis, The University of Sydney, 2008.

L. Song and B. Dai. Robust low rank kernel embeddings of multivariate distributions. In *Advances in Neural Information Processing Systems 26*, pages 3228–3236. Curran Associates Inc., 2013.

L. Song, A. Smola, A. Gretton, and K. Borgwardt. A dependence maximization view of clustering. In *Proceedings of the 24th International Conference on Machine Learning*, pages 815–822. ACM, 2007a.

L. Song, A. Smola, A. Gretton, K. Borgwardt, and J. Bedo. Supervised feature selection via dependence estimation. In *Proceedings of the 24th International Conference on Machine Learning*, pages 823–830. ACM, 2007b.

L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the 25th International Conference on Machine Learning*, pages 992–999. ACM, 2008.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th International Conference on Machine Learning*, pages 961–968. ACM, 2009.

L. Song, B. Boots, S. Siddiqi, G. Gordon, and A. Smola. Hilbert space embeddings of hidden Markov models. In *Proceedings of the 27th International Conference on Machine Learning*, pages 991–998. Omnipress, 2010a.

L. Song, A. Gretton, and C. Guestrin. Nonparametric tree graphical models via kernel embeddings. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, volume 9, pages 765–772, 2010b.

L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel belief propagation. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 707–715, 2011a.

L. Song, A. Parikh, and E. Xing. Kernel embeddings of latent tree graphical models. In *Advances in Neural Information Processing Systems*, pages 2708–2716, 2011b.

L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search.* MIT Press, 2000.

M. Springer. *The Algebra of Random Variables.* John Wiley & Sons, 1979.

B. Sriperumbudur. Mixture density estimation via Hilbert space embedding of measures. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 1027–1030, 2011.

B. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

B. Sriperumbudur and Z. Szabó. Optimal rates for random Fourier features. In *Advances in Neural Information Processing Systems 28*, pages 1144–1152. Curran Associates, Inc., 2015.

B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *The 21st Annual Conference on Learning Theory*, pages 111–122. Omnipress, 2008.

B. Sriperumbudur, K. Fukumizu, A. Gretton, G. Lanckriet, and B. Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems 22*, pages 1750–1758. Curran Associates Inc., 2009.

B. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 99:1517–1561, 2010.

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011a.

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Learning in Hilbert vs. Banach spaces: A measure embedding viewpoint. In *Advances in Neural Information Processing Systems 24*, pages 1773–1781. Curran Associates, Inc., 2011b.

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.

B. Sriperumbudur, K. Fukumizu, A. Gretton, and A. Hyvärinen. Density estimation in infinite dimensional exponential families. *CoRR*, abs/:1312.3516, 2013.

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206. University of California Press, 1955.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2002.

I. Steinwart and A. Christmann. *Support Vector Machines.* Springer Publishing Company, Incorporated, 1st edition, 2008.

I. Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

H. Strathmann, D. Sejdinovic, S. Livingstone, Z. Szabó, and A. Gretton. Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families. In *Advances in Neural Information Processing Systems 28*, pages 955–963. Curran Associates, Inc., 2015.

L. Su and H. White. A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, 24:829–864, 2008.

E. Sudderth, A. Ihler, M. Isard, W. Freeman, and A. Willsky. Nonparametric belief propagation. *Communications of the ACM*, 53(10):95–103, 2010.

D. Sutherland and J. Schneider. On the error of random Fourier features. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 862–871. AUAI Press, 2015.

R. Sutton and A. Barto. *Introduction to Reinforcement Learning*. MIT Press, 1st edition, 1998.

Z. Szabó, B. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152): 1–40, 2016.

G. Székely and M. Rizzo. Testing for equal distributions in high dimensions. *InterStat*, 2004.

G. Székely and M. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005.

G. Székely and M. Rizzo. Brownian distance covariance. *Annals of Applied Statistics*, 3(4):1236–1265, 12 2009.

G. Székely, M. Rizzo, and N. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, 2007.

I. Tolstikhin, B. Sriperumbudur, and K. Muandet. Minimax estimation of kernel mean embeddings. *CoRR*, abs/:1602.04361, 2016.

I. Tsamardinos and G. Borboudakis. Permutation testing improves Bayesian network learning. In *ECML/PKDD (3)*, volume 6323 of *Lecture Notes in Computer Science*, pages 322–337. Springer, 2010.

A. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

V. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer Verlag, New York, 2nd edition, 2000.

A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):480–492, 2012.

L. Wasserman. *All of Nonparametric Statistics*. Springer-Verlag New York, Inc., 2006.

L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated, 2010.

M. Welling. Herding dynamical weights to learn. In *Proceedings of the 26th International Conference on Machine Learning*, pages 1121–1128. ACM, 2009a.

M. Welling. Herding dynamic weights for partially observed random field models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 599–606. AUAI Press, 2009b.

M. Welling and Y. Chen. Statistical inference using weak chaos and infinite memory. *Journal of Physics: Conference Series*, 233(1), 2010.

H. Wendland. *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK, 2005.

C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.

S. Wu and S.-I. Amari. Conformal transformation of kernel functions: A data-dependent way to improve support vector machine classifiers. *Neural Processing Letters*, 15(1):59–67, 2002.

L. Xiong, B. Poczos, and J. Schneider. Group anomaly detection using flexible genre models. In *Advances in Neural Information Processing Systems 24*, pages 1071–1079. Curran Associates Inc., 2011a.

L. Xiong, B. Póczos, J. Schneider, A. Connolly, and J. VanderPlas. Hierarchical probabilistic models for group anomaly detection. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15, pages 789–797. JMLR.org, 2011b.

J. Yang, V. Sindhwani, H. Avron, and M. Mahoney. Quasi-Monte Carlo feature maps for shift-invariant kernels. In *Proceedings of the 31th International Conference on Machine Learning*, volume 32, pages 485–493. JMLR.org, 2014.

Y. Yoshikawa, T. Iwata, and H. Sawada. Latent support measure machines for bag-of-words data classification. In *Advances in Neural Information Processing Systems 27*, pages 1961–1969. Curran Associates, Inc., 2014.

Y. Yoshikawa, T. Iwata, and H. Sawada. Non-linear regression for bag-of-words data via Gaussian process latent variable set model. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 3129–3135, 2015.

W. Zaremba, A. Gretton, and M. Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in Neural Information Processing Systems 26*, pages 755–763. Curran Associates, Inc., 2013.

H. Zhang, Y. Xu, and J. Zhang. Reproducing kernel Banach spaces for machine learning. *Journal of Machine Learning Research*, 10:2741–2775, 2009.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 804–813. AUAI Press, 2011.

X. Zhang, L. Song, A. Gretton, and A. Smola. Kernel measures of independence for Non-IID data. In *Advances in Neural Information Processing Systems 21*, pages 1937–1944. Curran Associates, Inc., 2008.

J. Zhao and D. Meng. FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *Neural Computation*, 27(6):1345–1372, 2015.

L. Zwald, O. Bousquet, and G. Blanchard. Statistical properties of kernel principal component analysis. In *Proceedings of the 17th Annual Conference on Learning Theory*, volume 3120 of *Lecture Notes in Computer Science*, pages 594–608. Springer, 2004.