Feature extraction and selection using statistical dependence criteria

Diego Tomassi, Nicolás Marx, and Pierre Beauseroy

[1] Instituto de Matemática Aplicada del Litoral,
UNL-CONICET-FIQ-FICH, Santa Fe, Argentina
[2] Université de Technologie de Troyes, France

Abstract. Dimensionality reduction using feature extraction and selection approaches is a common stage of many regression and classification tasks. In recent years there have been significant efforts to reduce the dimension of the feature space without lossing information that is relevant for prediction. This objective can be cast into a conditional independence condition between the response or class labels and the transformed features. Building on this, in this work we use measures of statistical dependence to estimate a lower-dimensional linear subspace of the features that retains the sufficient information. Unlike likelihood-based and many moment-based methods, the proposed approach is semi-parametric and does not require model assumptions on the data. A regularized version to achieve simultaneous variable selection is presented too. Experiments with simulated data show that the performance of the proposed method compares favorably to well-known linear dimension reduction techniques.

Keywords: Dimension reduction; variable selection; dependence measures; supervised learning.

1 Introduction

In a supervised learning scenario, we are often interested in building a predictive rule for a response variable Y, which can be discrete or continuous, according to some set of covariates $\mathbf{X} \in \mathbb{R}^p$. When p is large, finding such a rule is challenging, since many of the covariates tipically do not carry relevant information about Y. On the contrary, they contribute to increase the variance of the estimates, which often affects the generalization ability of the predictive machine. Motivated by this, dimensionality reduction has long been a basic problem for many machine learning and pattern recognition applications. Two tasks are frequently distinguished: dimension reduction as a feature extraction process and variable selection. Though the LASSO-family of methods [1] represents currently a well-stablished methodology for variable selection with generalized linear predictive models, there is not a method recognized widely as the gold-standard for supervised dimension reduction. In recent years, a methodology called *sufficient dimension reduction* (SDR) has attracted interest both in the statistics and machine learning communities [2,7]. The distinctive attribute of SDR is to formally take

care about preserving the information about the response when searching for a lower-dimensional subspace of the features for prediction. This objective can be formalized as a conditional independence statement involving the response and the transformed features. Building on this, in this work we use measures of statistical dependence to drive the estimation of the reduction. This approach does not rely on particular assumptions on $Y|\mathbf{X}$, $\mathbf{X}|Y$ or on the marginal distribution of $\mathbf{X}$. The proposed method proceeds sequentially, extracting one direction of projection at each iteration. Moreover, a regularized version of the estimator using a mixed-norm penalty is presented, allowing for simultaneous variable selection without needing to assume any predictive rule.

1.1 Related work

Masaeli et al [14] proposed a simultaneous feature extraction and selection method based on maximizing a non-conditional generalized measure of correlation between the response and the reduced features. A penalty term in their approach also allows for goup-lasso-type regularization. Nevertheless, their approach do not take care of information preservation and maximizing correlation can favour specific types of dependence. Semi-parametric kernel-based methods for dimension reduction in regression were proposed in [7]. The proposed algorithm, however, requires several inversions of large Gram matrices, which become computationally infeasible when the dimension of the features increases. Fukumizu and Leng [9] proposed a linear dimension reduction method based on an estimate of the gradient of the regression function for feature vectors mapped to reproducing kernel Hilbert Spaces (RKHS). Despite it relaxes the computational cost of [7], no insight is provided about the dimension of the smallest subspace that retains the relevant information.

2 Background

2.1 Sufficient dimension reduction

Sufficient dimension reduction (SDR) is a methodology that aims at finding a lower dimensional subspace of the original features that retains all the information about the response [2,12]. SDR methods mostly look for optimal *linear transformations* of the predictors. Let $\boldsymbol{\beta}$ be a semi-orthogonal basis matrix for the lower-dimensional subspace and let $F(A|\cdot)$ indicate the conditional distribution function of A given the second argument. The transformed feature vector $\boldsymbol{\beta}^T\mathbf{X} \in \mathbb{R}^d$, with $d \leq p$, is a sufficient reduction to predict Y if

$$F(Y \mid \mathbf{X}) = F(Y \mid \boldsymbol{\beta}^T\mathbf{X}). \tag{1}$$

The parameter of interest is actually the subspace spanned by the columns of $\boldsymbol{\beta}$, not $\boldsymbol{\beta}$ itself, and the goal in SDR is to find the smallest subspace where (1) holds.

Estimation in SDR is commonly carried out using the inverse regression $\mathbf{X}|Y$. A first approach under this setting is to consider functions of moments of $\mathbf{X}|Y$

to drive the estimation. Some examples include Sliced Inverse Regression (SIR) [12], Sliced Average Variance Estimation (SAVE) [4] and Directional Regression (DR) [11]. Despite being easy to compute, these methods require different conditions on the marginal distribution of $\mathbf{X}$ to yield sufficient reductions. Moreover, obtained reductions are not necessarily exhaustive, in the sense that they might not include all the functions of the covariates needed to describe the response.

Another approach under the inverse regression framework is to consider likelihood-based estimation, as first considered in [2]. If the distribution of $\mathbf{X}|Y$ is available, maximum-likelihood estimates of the reduction can be derived from the fact that

$$F(Y \mid \mathbf{X}) = F(Y \mid \boldsymbol{\beta}^T\mathbf{X}) \Leftrightarrow F(\mathbf{X} \mid \boldsymbol{\beta}^T\mathbf{X}, Y) = F(\mathbf{X} \mid \boldsymbol{\beta}^T\mathbf{X}). \tag{2}$$

Unlike moment-based methods, likelihood-based estimation guarantees to estimate the whole sufficient reduction when the model assumptions hold. Nevertheless, checking distribution assumptions on the data is often a non-trivial task, and optimality of the methods for a particular problem becomes hard to assess.

A third approach to drive the estimation of the reduction relies on the following statement which also characterizes a sufficient reduction: $\boldsymbol{\beta}^T\mathbf{X}$ is sufficient to predict Y if

$$F(Y, \mathbf{X} \mid \boldsymbol{\beta}^T\mathbf{X}) = F(Y \mid \boldsymbol{\beta}^T\mathbf{X})F(\mathbf{X} \mid \boldsymbol{\beta}^T\mathbf{X}). \tag{3}$$

Solutions based on this criterion are currently limited to kernel-based methods using cross-covariance operators in reproducing kernel Hilbert spaces (RKHS) [7,9]. We follow this line in this work.

3 Proposed method

Let $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ be a semi-orthogonal basis matrix ($\boldsymbol{\beta}^T\boldsymbol{\beta} = \mathbf{I}_d$) for the subspace $\mathcal{S}_\beta$ and let $\boldsymbol{\beta}_0 \in \mathbb{R}^{p\times(p-d)}$ be a basis matrix for the orthogonal complementary subspace, with $\boldsymbol{\beta}_0^T\boldsymbol{\beta}_0 = \mathbf{I}_{p-d}$, $\boldsymbol{\beta}^T\boldsymbol{\beta}_0 = \mathbf{0}$ and $\boldsymbol{\beta}_0^T\boldsymbol{\beta} = \mathbf{0}$. Clearly, condition (3) for SDR implies

$$(Y \perp\!\!\!\perp \boldsymbol{\beta}_0^T\mathbf{X}) \mid \boldsymbol{\beta}^T\mathbf{X}, \tag{4}$$

where $\perp\!\!\!\perp$ indicates statistical independence. Let $\gamma^2(U, V|Z)$ be a generalized measure of conditional dependence between random vectors U and V, given Z, so that $\gamma^2(U, V|Z) = 0$ if and only if $(U \perp\!\!\!\perp V)|Z$. Assuming that dimension d is known, we can estimate $\boldsymbol{\beta}$ just solving

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}:\, \boldsymbol{\beta}^T\boldsymbol{\beta}=\mathbf{I}}{\arg\min} \left\{ \gamma(Y, \boldsymbol{\beta}_0^T\mathbf{X} \mid \boldsymbol{\beta}^T\mathbf{X}) + \lambda \sum_{j=1}^{p} \|\boldsymbol{\beta}_j\|_2 \right\}, \tag{5}$$

where $\boldsymbol{\beta}_j$ is the j-th row of $\boldsymbol{\beta}$. The penalty term aims at pushing some rows of $\widehat{\boldsymbol{\beta}}$ to zero, meaning that only a subset of the features are retained in the final estimate.

Algorithm 1 Sequential estimation of $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p \times d_0}$

Inputs: $(\mathbb{X}, \mathbb{Y}, d_0, \lambda)$;
Outputs: $\hat{\boldsymbol{\beta}}$
procedure
  Set $\mathbf{W} = \emptyset$, $\mathbf{W}_0 = \mathbf{I}_p$
  for $j = 1$ to d_0 **do**
    $\mathbb{Z} = \mathbb{X}\mathbf{W}_0$
    $\tilde{\boldsymbol{\beta}}_{\text{ini}} \leftarrow$ Compute initial estimate from $(\mathbb{Z}, \mathbb{Y})$
    Using $\tilde{\boldsymbol{\beta}}_{\text{ini}}$, compute

$$\tilde{\boldsymbol{\beta}}_{\text{aux}} = \underset{\boldsymbol{\beta}:\ \boldsymbol{\beta}^T\boldsymbol{\beta}=\mathbf{I}}{\arg\min}\ \rho_n(\mathbb{Z}\boldsymbol{\beta}_0, \mathbb{Y}|\mathbb{Z}\boldsymbol{\beta}) + \lambda \sum_j ||\tilde{\boldsymbol{\beta}}_j||_2$$

    $\tilde{\boldsymbol{\beta}}^{(j)} = \mathbf{W}_0\boldsymbol{\beta}_{\text{aux}}$
    $\mathbf{W} = (\tilde{\boldsymbol{\beta}}^{(1)}, \ldots, \tilde{\boldsymbol{\beta}}^{(j)})$
    $\mathbf{W}_0 \leftarrow$ take orthogonal complement of $\mathbf{W}$
  end for
  for $j = 1$ to p **do**
    if $||\mathbf{w}_j||_2 < Thr$ **then**
      set row j to zero
    end if
  end for
  $\hat{\boldsymbol{\beta}} = \mathbf{W}$
end procedure

Instead of direct estimation using (5), we propose to compute $\boldsymbol{\beta}$ sequentially, obtaining its columns iteratively one-by-one. The proposed sequential procedure is outlined in Algorithm 1. The motivation to proceed in this way is two-fold. On the one hand, sequential estimation implies that the conditioning in (5) is one-dimensional, which is often easier to compute. On the other hand, we can take advantage of the sequential procedure to automatically choose the dimension d. The rationale is as follows. Assume that the true, unknown dimension is d_0. After iteration k, the value of the dependence measure will be greater than zero for $k < d_0$, since $\boldsymbol{\beta}_0^T\mathbf{X}$ still carries some information about Y. Nevertheless, for $k \geq d_0$, the subspace spanned by $\boldsymbol{\beta}_0$ no longer has information about Y and the population dependence measure will be zero. For dependence measures with good convergence properties, we expect to observe a similar behaviour in their sample estimates. Thus, we can track the value of the dependence measure at each iteration to decide when to stop searching for an additional direction of projection.

A key aspect of the proposed method is the availability of a suitable measure of statistical dependence. Several such measures have been proposed in the literature recently, mainly driven by the interest in finding associations between covariates in large data sets (a hypothesis testing problem) [8,10,13,15]. Due

to space constraints, in this work we restrict the discussion and results to a conditional version of a Hilbert-Schmidt norm between covariance operators in RKHS. See [8] for details. This is related to the criteria used in [7,9,14] and allows for a more direct comparison with them. Finally, the algorithm requires to provide suitable starting values. Here we use a simpler estimate for the direction of projection, known as approximate information discriminant analysis (AIDA) [5] which is computed from an eigenanalysis of a symmetric matrix.

4 Experiments

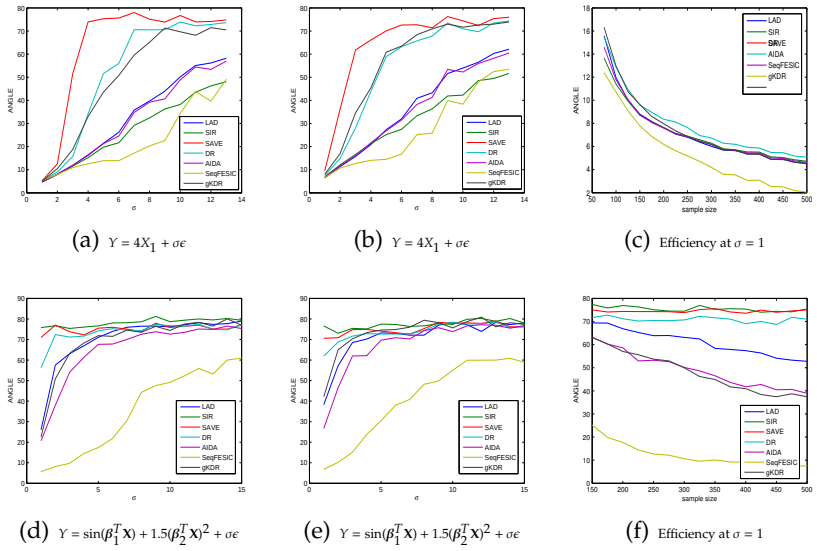
4.1 Accuracy and efficiency of estimation

The following data models are studied as examples:

i) $Y = 4X_1 + \sigma\epsilon$,
ii) $Y = \sin(\boldsymbol{\beta}_1^T\mathbf{X}) + 1.5(\boldsymbol{\beta}_2^T\mathbf{X})^2 + \sigma\epsilon$,

Case i) provides the simplest example of linear model, where only one variable is relevant to predict the response. On the other hand, case iii) presents a model where all the original features are active to describe the response, but on a two-dimensional characteristic subspace. In addition, for both cases i) and iii), two conditions are evaluated: a) $\epsilon \sim \mathcal{N}(0,1)$; and b) ϵ is distributed as a mixture $\epsilon \sim 0.3t_4 + 0.7\chi^2_{(2)}$. For all cases, we set $\mathbf{X} \in \mathbb{R}^{10}$, with $X_j \sim \mathcal{N}(0,1)$ for all j. For case iii), $\boldsymbol{\beta} = (\boldsymbol{\beta}_1\boldsymbol{\beta}_2)$ is generated at random and held fixed for the experiment. Obtained performance is compared with other linear dimension reduction methods, like SIR [12], SAVE [4], DR [11], LAD [3], AIDA [5], and gKDR [9]. We refer to the proposed algorithm as SeqFESIC (which stands for Sequential Feature Extraction and Selection using Independence Criteria). Accuracy of estimation is measured in terms of the angle θ between the true subspace and the estimated one; that is, $\theta = \text{angle}(\mathcal{S}_{\boldsymbol{\beta}} - \mathcal{S}_{\widehat{\boldsymbol{\beta}}})$ (see [6] for details).

Figure 4.1 shows the obtained results for the quality of estimation as a function of σ, using a training sample of $n = 500$ points. Subfigures a) and c) correspond to the case of Gaussian noise, while subfigures b) and d) correspond to the non-normal condition. Reported results are averages over 100 runs of the experiment. It can be seen that for case i) with normal noise, SeqFESIC achieves results very similar to those of LAD, SIR and AIDA. Note that LAD is an optimal estimator under this setting and that it is also very favorable to SIR. The proposed method obtains slightly larger angles when σ is small, but it obtains better results than the competing methods when the noise becomes stronger. Note also that SeqFESIC shows better results than gKDR over all the range of σ. Switching to the results for the non-normal noise condition for case i), it can be seen that performance for all the methods remain close to those otained under Gaussian noise. Finally, for the more general setting described in case ii), the proposed method is clearly superior than all the other techniques for all the values of σ, both with Gaussian noise and with non-Gaussian noise.



(a) $Y = 4X_1 + \sigma\epsilon$ (b) $Y = 4X_1 + \sigma\epsilon$ (c) Efficiency at $\sigma = 1$

(d) $Y = \sin(\beta_1^T \mathbf{X}) + 1.5(\beta_2^T \mathbf{X})^2 + \sigma\epsilon$ (e) $Y = \sin(\beta_1^T \mathbf{X}) + 1.5(\beta_2^T \mathbf{X})^2 + \sigma\epsilon$ (f) Efficiency at $\sigma = 1$

Fig. 1. Accuracy of estimation, measured with the angle between span($\boldsymbol{\beta}$) and span($\widehat{\boldsymbol{\beta}}$). (a) and (d) correspond to normal noise ϵ, while (b) and (e) correspond to a non-normal noise condition. (c) and (f) show the quality of the estimation as a function of n.

We further studied how the sample size of the training set affects the estimation quality of the proposed method. The same data models from Section 4.1 were used, with zero-mean normal errors and variance $\sigma^2 = 2^2$. Note that for this noise condition, the performance of all methods is similar for $n = 500$. Here we let the size of the training sample change between $n = 75$ and $n = 500$. Obtained results are shown in panels (c) and (e) of Figure 4.1 for case i) and case ii), respectively. Reported values are averages over 100 runs of the experiment. It can be seen that in both scenarios SeqFESIC obtains the best scores for smaller sample sizes. It is followed by SIR, LAD and AIDA for data as in case i), and by gKDR and AIDA for data as in case ii). A main reason for these results is that SeqFESIC does not require covariance matrix estimators for its computation, which is indeed the case for all the other methods with the exception of gKDR.

4.2 Accuracy of variable selection

In some applications, regularization not only aims at obtaining better scores in prediction, but also to identify a subset of the original features that actually do not contribute to explain the response. In this section we study if the proposed algorithm achieves this goal. Similarly to Section 4.1, the following linear model is used:

$$Y = \sin(\boldsymbol{\beta}_1^T \mathbf{X}) + 1.5(\boldsymbol{\beta}_2^T \mathbf{X})^2 + \epsilon.$$

Table 1. Performance for variable selection

	DCOR($\mathbf{X}_A, \mathbf{X}_I$)									
	0		0.15		0.30		0.45		0.65	
	SeqFESIC	HSFS	SeqFESIC	HSFS	SeqFESIC	HSFS	SeqFESIC	HSFS	SeqFESIC	HSFS
$\hat{r}_1$:	.986	.982	.974	.968	.958	.922	.912	.834	.876	.722
$\hat{r}_2$:	.062	.060	.092	.084	.244	.360	.410	.680	.604	.786
$\hat{r}_3$:	4.12	4.15	4.47	4.35	5.31	5.89	5.88	8.54	7.19	10.63

We set $\mathbf{X} \in \mathbb{R}^{100}$, with $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, $\boldsymbol{\beta}^T = (\tilde{\boldsymbol{\beta}}_A^T \ \mathbf{0})$, $\tilde{\boldsymbol{\beta}}_A^T = (\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T)$, $\tilde{\boldsymbol{\beta}}_1^T = (1\ 1\ 1\ 1)/2$, $\tilde{\boldsymbol{\beta}}_2^T = (-1\ 0\ 0\ 1)/\sqrt{2}$ and $\epsilon \sim \mathcal{N}(0, 1)$. In this way, $Y = g(\mathbf{X}_A, \mathbf{X}_I) + \epsilon$, with $\mathbf{X}_A \in \mathbb{R}^4$ and $\mathbf{X}_I \in \mathbb{R}^{96}$. To control the amount of correlation, we modify $\mathbf{\Sigma}$ and use distance correlation [15] between the active and nonrelevant set of features, DCOR($\mathbf{X}_A, \mathbf{X}_I$), as a measure of interaction between the relevant and non-relevant predictors. For a specified value of DCOR($\mathbf{X}_A, \mathbf{X}_I$), a realization is discarded if the sample statistic of DCOR differs more than ±5% from the target value. Results obtained with the proposed algorithm are compared to those obtained with the method introduced in [14], which will be refered to as HSFS (Hilbert-Schmidt Feature Selection).

Let S_0 be the set of indices of the relevant predictors, and let $\hat{S}$ be the subset of variables selected as relevant to predict the response Y. Performance of the variable selection procedure is assessed using the following criteria: $r_1 \equiv \mathrm{p}(\hat{S} \supseteq S_0)$, as an indicative of the fraction of times that the relevant predictors are preserved; $r_2 \equiv \mathrm{p}(\hat{S}^c \subset S_0^c)$, as an indicative of the fraction of times that nonrelevant predictors are retained along with the relevant ones; and $r_3 \equiv \mathrm{E}(Card_{\hat{S}})$, with $Card_{\hat{S}} = \#\{\hat{S} | \hat{S}^c \subset S_0^c\}$, as a measure of the average number of nonrelevant features that are retained along with the relevant ones.

Obtained results are reported in Table 1. Quantities were estimated after 500 runs of the experiment. It can be seen that when the correlation between the active set of predictors and the irrelevant ones is low, the variable selection procedure is very accurate in picking the true active predictors. For DCOR($\mathbf{X}_A, \mathbf{X}_I$) ≤ 0.15, the average number of retained variables is less than 5 (only one more than the true value 4). Most important, among the retained ones, more than 97% of the times this retained subset contains all the relevant predictors. For larger values of DCOR($\mathbf{X}_A, \mathbf{X}_I$) performance degrades, as it is expected. Nevertheless, even for higher values of correlation up to 0.45, the retained set of predictors contains the true one more than 90% of the runs. Values for $\hat{r}_2$ suggest that under these conditions 40% of the times the algorithm picks some irrelevant features, but the total number of chosen predictors remains less than 6, as indicated by $\hat{r}_3$. On the other hand, HSFS gets similar scores for scenarios of mild correlation, but its performance degrades faster when correlation increases. For DCOR($\mathbf{X}_A, \mathbf{X}_I$) ≥ 0.45, HSFS retains significantly more variables than SeqFESIC,while, at the same time, it fails to pick the relevant ones more than 17% of the runs.

5 Conclusion

A new method for feature extraction and variable selection using measures of statistical dependence was introduced. Results with simulated data show that it compares favorably to well-known sufficient dimension reduction methods, even under experimental settings where some of those methods are optimal. Moreover, this difference in performance seems to be more important in noisy scenarios or with very limited data. Further experiments with large-scale real datasets are needed to confirm these results and to assess the scalability of the method to real-world applications.

References

1. Peter Buhlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, New York, 2011.
2. R.D. Cook. Fisher lecture: Dimension reduction in regression (with discussion). *Statistical Science*, 22:1–26, 2007.
3. R.D. Cook and L. Forzani. Likelihood-Based sufficient dimension reduction. *Journal of the American Statistical Association*, 104(485):197–208, 2008.
4. RD Cook and S. Weisberg. Discussion of sliced inverse regression. *Journal of the American Statistical Association*, 86:328–332, 1991.
5. K. Das and Z. Nenadic. Approximate information discriminant analysis: A computationally simple heteroscedastic feature extraction technique. *Pattern Recognition*, 41(5):1565–1574, 2008.
6. M. Eaton. *Multivariate Statistics*. Wiley, New York, 1983.
7. 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.
8. K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in neural information processing systems*, pages 489–496, 2008.
9. K. Fukumizu and C. Leng. Gradient-based kernel dimension reduction for regression. *Journal of the American Statistical Association*, 109(505):359–370, 2014.
10. 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*, ALT'05, pages 63–77, 2005.
11. Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.
12. K.C. Li. Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86:316–342, 1991.
13. D. Lopez-Paz, P. Hennig, and B. Schölkopf. The randomized dependence coefficient. In *Advances in Neural Information Processing Systems*, pages 1–9. 2013.
14. M. Masaeli, J. Dy, and G. Fung. From transformation-based dimensionality reduction to feature selection. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 751–758, 2010.
15. Gabor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.