INDEPENDENT COMPONENT ANALYSIS WITH QUANTIZING DENSITY ESTIMATORS

Peter Meinicke, Helge Ritter

Neuroinformatics Group
University Bielefeld
Germany

ABSTRACT

We propose an approach to source adaptivity in ICA based on quantizing density estimators (QDE). These estimators allow to realize source adaptivity in an efficient and non-parametric way and we show how their application can be viewed as a natural extension to recent approaches based on parametric models. In simulations we show that ICA based on QDE can considerably increase the performance of blind source separation as compared with flexible parametric approaches.

1. INTRODUCTION

As shown by many authors [1, 2, 3, 4, 5, 6], independent component analysis (ICA) can be conveniently stated as a maximum likelihood (ML) estimation problem, which requires to recover a linear transformation of an observable random vector with the resulting components being statistically as independent as possible. In principle, to find the ML solution the original source distributions have to be known. Because in most real world situations this knowledge does not exist the model building process often involves some degree of arbitrariness. However, as indicated in [7], ICA can be realized without knowing the source distributions, since these distributions may be learnt from the data, together with the linear transformation. In that way, asymptotically[1] one may obtain the same performance in blind source separation (BSS), as if the distributions were known.

Recently a non-parametric approach to source adaptivity has been proposed [8] which utilizes the kernel density estimator (KDE), one of the most common methods for non-parametric density estimation [9, 10]. However some difficulties arise as one utilizes the KDE for source adaptivity in ICA and we shall indicate two problems in the following.

One problem arises if one tries to normalize the component distributions to unit variance. The assumption of unit variance sources is quite common in latent variable modelling and in particular in ICA, since it reflects the fact that the components can only be recovered up to an inherent scale indeterminacy. In ICA a suitable component normalization allows to realize learning via an unconstrained optimization of the separating matrix which gives rise to the common gradient descent schemes. Thus in the following we consider a component as a zero-mean random variable y with unit variance. Based on a sample $\{y_1, \ldots, y_N\}$ we have the kernel density estimator

$$\hat{f}(y) = \frac{1}{N} \sum_{i=1}^{N} K(y, y_i) \tag{1}$$

which for normalized symmetric kernel functions $K(y, \cdot)$ implies the following variance of the associated component distribution

$$\int_{-\infty}^{\infty} y^2 \hat{f}(y) dy = \int_{-\infty}^{\infty} y^2 K(y, 0) dy + \frac{1}{N} \sum_{i=1}^{N} y_i^2. \tag{2}$$

While the first (integral) term equals the variance of a zero-mean random variable distributed according to the kernel (density) function the second term equals the sample variance, which is an estimator of the component variance. Thus for a fixed non-zero kernel bandwidth we will have a systematic mismatch between the predicted variance as implied by the KDE and the direct estimator of the component variance. If we normalize the sample to have unit variance, then this deviation will asymptotically vanish since the kernel bandwidth is required to converge to zero for $N \to \infty$. Note that the component normalization to a fixed sample variance is in no way optional but necessary, because otherwise within a maximum likelihood setting an unconstrained ICA learning scheme utilizing the KDE will increase that variance[2] without bounds to maximize the likelihood of the model. With an appropriate rescaling of the components during optimization, in the asymptotic limit we will have a unit variance component, but for finite samples which always require non-zero bandwidths the inherent bias as introduced by the KDE may be large. Fortunately, this type of bias is not a necessary shortcoming of non-parametric

[1] for infinite data

[2] by increasing the norm of the separating matrix

estimation and in the next section we will suggest an alternative approach based on quantizing density estimators (QDE) [11].

Another problem with the KDE is the high computational cost for evaluation of the associated density function. With an N-point sample, ICA batch learning requires the (repeated) evaluation of the KDE at N distinct locations which results in a total complexity $O(N^2)$. For blind source separation, ICA is typically applied to large data sets and in these cases the computational complexity of the KDE would be prohibitive. With a technique originally suggested in [12], the authors of [8] therefore proposed an approximation of the KDE by means of a smoothed histogram. Using the Fast Fourier Transform (FFT) the complexity of N-point evaluation could be reduced to roughly $O(N \log N)$. In contrast to the KDE the computational cost for evaluation of the QDE does not directly depend on the data set size and in general the required complexity for N-point evaluation is far below quadratic. In the following section we shall now go into detail and the general form of the QDE will be introduced together with a specialization well-suited to address the source adaptivity issues in ICA.

2. QUANTIZING DENSITY ESTIMATORS

Quantizing density estimators (QDE) have recently been proposed as a general method for unsupervised learning [11], which can realize complexity-reduced representations of the data in a non-parametric way. Traditional techniques for dimensionality reduction and clustering, like PCA or vector quantization are well-suited for implementation with QDE. An important feature of the QDE is, that by the specification of a suitable quantizer one may include domain knowledge or algorithmic constraints without resorting to the strong distributional assumptions of the Bayesian framework.

In the following we will explain how the QDE may be derived from a generalization of the traditional KDE. Consider the situation where the KDE of (1) is constructed on the basis of a quantized sample. We then have the following estimator in 1D space

$$\hat{f}(y) = \frac{1}{N} \sum_{i=1}^{N} K(y, q(y_i; \boldsymbol{\theta})) \tag{3}$$

where $q : \mathbb{R} \to \mathcal{P} \subseteq \mathbb{R}$ is a given quantization or projection function which maps a point to a parametrized subset of the sample space according to

$$q(y; \boldsymbol{\theta}) = p(s_p(y); \boldsymbol{\theta})$$

with the projection index

$$s_p(y) = \arg\min_{z \in \mathcal{Z}} |y - p(z; \boldsymbol{\theta})|.$$

The projection index associates a data point with its nearest neighbour in the projection set

$$\mathcal{P} = \{y : y = p(z; \boldsymbol{\theta}),\ z \in \mathcal{Z}, \boldsymbol{\theta} \in \Theta\} \tag{4}$$

where $\mathcal{Z} \subseteq \mathbb{R}$ is the set of all possible projection indices. For an intuitive motivation of the QDE, one may ask from a data compression perspective whether it is necessary to store all the sample data $\{y_1, \ldots, y_N\}$ for the realization of the kernel density estimator or if it is possible to first reduce the data by some suitable quantization method and then construct the estimator from the more parsimonious complexity-reduced data set without decreasing its performance.

In one dimension a natural projection set can be specified by a set of M quantization levels on the real line, i.e. $\mathcal{P} = \{w_1, \ldots, w_M\}$. In order to minimize the Kullback-Leibler divergence between the model and the true distribution we can now perform maximum likelihood estimation of the level coordinates. In that way we obtain a maximum likelihood estimator which takes the form of a constrained mixture density

$$\hat{f}(y) = \frac{1}{N} \sum_{i=1}^{M} n_i K(y, \hat{w}_i) \tag{5}$$

with $n_i = |\{j : i = \arg\min_k |y_j - \hat{w}_k|\}|$ counting the number of data points which are quantized to level i.

From a different starting point the authors of [13] derived the same functional form of a non-parametric ML density estimator which they proposed as a variation of the KDE. In contrast to the KDE with fixed kernels centered on N data points the locations of the kernels were considered as parameters. As with the traditional KDE, for consistency of the estimator the bandwidth has to be decreased as the sample size increases. The authors in [13] reported that for a fixed non-zero kernel bandwidth ML-estimation of the N kernel centers always resulted in a smaller number of actually distinct centers, i.e. several kernels coincided to maximize the likelihood. Therefore the resulting estimator had the form of (5) where M corresponds to the number of distinct centers with n_i counting the number of kernels coinciding at w_i. The optimum number of effective quantization levels for a given bandwidth therefore arises as an automatic byproduct of the ML estimation of an N-point projection set $\mathcal{P} = \{w_1, \ldots, w_N\}$. Vice versa, as also shown in [13], for a fixed number of $M < N$ kernels the bandwidth may be estimated from the sample. In this case M has to be carefully increased with the sample size in order to guarantee consistency.

Within the ICA setting in essence we require some one-dimensional QDE which can be calculated in an efficient way. Therefore the above QDE (5) would not be the best choice since it introduces a considerable amount of computational cost, due to the high variability of the M-level

quantizer. For ML estimation of the parameters one has to alternate the projection of the data points to the M level positions (N nearest neighbour calculations) and the reestimation of these locations (e.g. via EM-algorithm). This iterative procedure has to be repeated each time when the separating matrix $\mathbf{W}$ has changed to some degree during the overall ICA learning scheme. To avoid bad local minima of the loss function it is recommendable to start the optimization with a small number of quantization levels and gradually increase the complexity of the quantizer. Such a model refinement requires additional effort and the overall computional cost become burdensome at least for large data sets. In the following we shall propose a suitable restriction on the above quantizer which leads to a simplified learning scheme with considerably reduced computing cost.

2.1. Regular Grids

A suitable simplification of the above QDE can be achieved by restricting the kernel centers to the nodes of a regular grid. For that purpose the projection set can be parametrized according to

$$p(z;\,\boldsymbol{\theta}) = \lambda(z-c), \quad z \in \{0,1,\ldots,M-1\} \tag{6}$$

where $c = (M-1)/2$ makes the grid symmetric w.r.t. the origin. Thus (6) yields an M-level quantizer which takes the form of a regular grid with distance λ between neighbouring nodes. The great simplification as compared with the previous quantizer is that now for a given number of quantization levels, the scale λ is the only parameter of the quantizer which needs to be estimated. In addition, projection onto that grid is trivial since it only requires an affine mapping and a simple rounding operation. For large M the FFT can be employed similar to [12] to further reduce the computational cost of QDE evaluation. With the above parametrization of the projection set the QDE takes the form of (5) with the kernel centers $\hat{w}_i$ replaced by $\hat{\lambda}(i-1-c)$.

Note, that with the above choice of a regular grid, symmetric distributions are favored, which reflects our expectation about most natural signal distributions. However even for distributions with high skewness the estimator will converge (at a slower rate) to the true density[3]. In that way also other "priors", which may for instance favor distributions with high kurtosis, may be implemented by a specific parametrization of the grid. For consistency the number of quantization levels is required to grow with the sample size such that the QDE converges to a mixture of delta functions.

[3]under some mild restrictions which are common in non-parametric density estimation (see e.g. [10])

3. LEARNING SCHEME

Within a zero-noise ML setting we seek a non-singular transformation $\mathbf{y} = \mathbf{W}\mathbf{x}$ of the observable random vector $\mathbf{x}$ with quadratic separating matrix $\mathbf{W}$. This gives rise to the following contrast or loss function

$$L(\mathbf{W},\boldsymbol{\theta}) = -n\log|\det\mathbf{W}| - \sum_{i=1}^{N}\sum_{j=1}^{d}\log f_j(y_{ij}) \tag{7}$$

which is the negative log-likelihood of the model. Thereby $f_j(y_{ij})$ denotes the j-th component density evaluated for dimension j of transformed data vector i. For an iterative minimization of (7) one may now alternate between optimizing $\mathbf{W}$ and $\boldsymbol{\theta}$ which contains the free density parameters. Minimization w.r.t. $\mathbf{W}$ can be achieved by gradient descent in terms of the natural [14] or relative [15] gradient, which both yield the following update rule for batch learning:

$$\mathbf{W} := \mathbf{W} - \eta(\mathsf{E}[\boldsymbol{\phi}(\mathbf{y})\mathbf{y}^T] - \mathbf{I})\mathbf{W} \tag{8}$$

with the expectation $\mathsf{E}[\cdot]$ estimated by the corresponding sample average. The vector of score functions $\boldsymbol{\phi}(\mathbf{y})$ has components according to

$$\phi_j(y) = -\frac{\partial \log f_j(y)}{\partial y} = -\frac{f_j'(y)}{f_j(y)} \tag{9}$$

Optimization of the QDE parameters is achieved by ML-estimation of the scale factors

$$\hat{\lambda}_j = \arg\max_{\lambda\in\Lambda}\sum_{i=1}^{N}\log f_j(y_{ij};\,\lambda). \tag{10}$$

For a simple 1D search we restrict Λ to a set of discrete values from a suitable interval. For each element from this set one has to perform a projection step which maps the y_{ij} to their nearest neighbours on the regular quantization grid. Then for each dimension we construct a QDE of the form

$$\hat{f}(y;\,\lambda) = \frac{1}{N}\sum_{i=1}^{M} n_i K(y, \lambda(i-1-c)) \tag{11}$$

and calculate the associated log-likelihood. Thereby the n_i count the number of points which have been quantized to $\lambda(i-1-c)$ and the kernel bandwidth is chosen to satisfy the unit variance constraint on the component distributions (see (13) for an example). In the following we will consider suitable kernel functions and propose an optimization heuristic for finding the global minimum of the loss function.

3.1. Choice of Kernels

We will now argue that at least for small data sets the ICA performance crucially depends on the type of kernel used for construction of QDE. For blind source separation it is important to find the global minimum of the in general highly non-convex loss function (7). For that purpose a well-known optimization heuristic is to start with a model of low complexity, which yields a more simple loss function which is at best convex or at least has a smaller number of local minima. The probability of reaching the global minimum is therefore higher than with more complex models and the idea is now to track an initial minimum over a sequence of model refinement steps which gradually deform the loss function. The practical results of such an optimization scheme are usually convincing (see [16] for an overview) and in the case of ICA with high source adaptivity we argue that such an approach may also greatly simplify the search for an appropriate optimum. With the above QDE it is therefore recommendable to start with single level quantizers which imply densities with a single kernel function and then to add further quantization levels as optimization proceeds. From this perspective it is clear that the success of an iterative model refinement method must depend on the choice of the kernel functions. With a Gaussian kernel the initial source densities would also be Gaussian and the global minimum of (7) can be found by decorrelation of the data variables. However, this optimum is usually far from an acceptable BSS solution and often a large amount of quantization levels is needed to find the desired solution.

For that reason we do not use Gaussian kernels and instead we propose to choose a suitable kernel from a set of candidate functions during a first ($M = 1$) optimization stage. Suitable candidate functions can be derived from the family of generalized Gaussian density functions, with Gaussian and Laplacian densities being special cases:

$$K(y,z) = \frac{a}{2h\Gamma(1/a)} \exp\left[-\left|\frac{y-z}{h}\right|^a\right] \qquad (12)$$

where $\Gamma(\cdot)$ is the Gamma function. Thereby the kernel width is specified by h while the shape is determined by a. For $a = 2$ we have a Gaussian density, for $a = 1$ we get the Laplacian density with positive kurtosis and for $a > 2$ the density becomes "flattened" with negative kurtosis. Therefore during the initial optimization stage with 1-level quantizers we estimate the kurtosis and switch to $a = 1$ for positive and to $a = 4$ for negative values. This scheme has been proposed as "flexible ICA" in [4] where the authors added a third $a = 0.8$ function to model sources with high kurtosis. This extension is not necessary with our QDE approach since higher kurtosis can be captured at later optimization stages with more quantization levels. Therefore after the first optimization stage the kernel shape is not changed anymore because the overall shape of the source density can now be adapted by the QDE based combination of kernel functions.

As already mentioned above, the QDE has to be repeatedly normalized to have unit variance. With the generalized Gaussian kernel for a certain value of the scale factor λ this is achieved by adjusting the bandwidth h to satisfy

$$h^2\frac{\Gamma(3/a)}{\Gamma(1/a)} + \frac{\lambda^2}{N}\sum_{j=1}^{M} n_j(j-1-c)^2 = 1. \qquad (13)$$

It is easy to see that this constraint also implies an upper bound on the scale factor λ.

4. EXPERIMENTS

To investigate the performance of the QDE based ICA algorithm as applied to blind source separation we generated four linear mixtures from four different sources, with associated density functions shown in figure 1. The sources exhibit several distributional characteristics, like kurtosis (positive: Laplacian, exponential; negative: uniform, mixture of two Gaussians), skewness (exponential) and bi-modality (mixture of two Gaussians). All densities correspond to unit-variance distributions. For better reproducibility we used the mixing matrix given in [4] to transform a 1000-point data set sampled from these sources. The above learning scheme was applied as follows: The initial separating matrix was set to a diagonal matrix with inverse (estimated) standard deviations of the corresponding data dimensions as entries. In the first optimization stage we started with one-level QDE with "switching" kernel shapes which corresponds to the flexible ICA algorithm in [4]. This stage was run for 200 iterations applying the gradient update rule (8) with learning rate $\eta = 0.05$. A kurtosis-dependent switching of the kernel shape was enabled after each 10-th iteration. The first stage was repeated for five times with slight random variations of the initial separating matrix and the matrix associated with the lowest value of the loss function (7) was used at the start of the following optimization stage. At the end of the first stage the kernel shape for each component was "freezed". The number M of quantization levels was doubled and the scale was selected from 20 candidate values on a regular grid within the interval $[0, 2]$ (with borders excluded). Starting with the smallest scale, increasing values were considered as long as the normalization constraint (13) could be satisfied. The scale with maximum log-likelihood was then selected for the subsequent optimization stage. After each stage M was doubled and the new scale was selected (as above) from the updated interval $[0, 2\lambda]$ with λ being the old scale. In addition the new learning rate was set to $\max\{0.5*\eta, 0.0005\}$ with η being the old rate. Then the subsequent optimization stage was again run for 200 iterations. After the 8-th stage with $M = 128$ the

optimization was aborted since convergence of the separating matrix was reached. The estimated component densities as shown in figure 3 indicate that more information couldn't be drawn from the 1000-point data set since one clearly sees that overfitting has already begun. For comparison also the resulting QDE of the 5-th stage with $M = 32$ are shown in figure 2 where the characteristics of the source distributions are already captured fairly well.

To monitor the separation performance we used Amari's error criterion (see e.g. [4]) which is plotted in figure 4 for every 10-th iteration. For comparison also the continuation of the flexible ICA scheme of the first stage has been monitored for the next 1400 iterations, with performance also plotted in figure 4. From the plot we see, that the flexible ICA continuation does not lead to further reduction of the error, whereas the QDE based continuation clearly improves the separation.

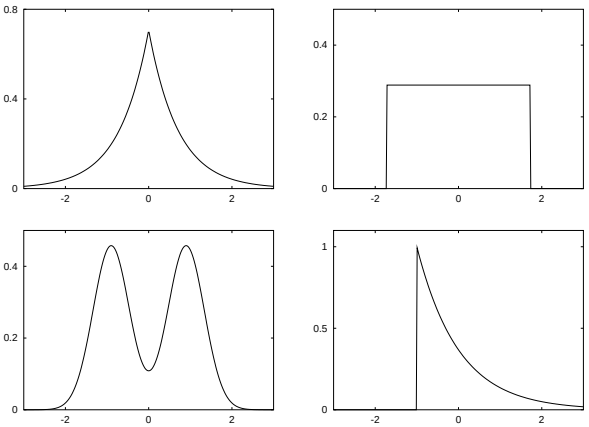

Fig. 1. Original source densities (from upper left to lower right): Laplacian, uniform, mixture of Gaussians and exponential density

5. CONCLUSION

We utilized quantizing density estimators (QDE) recently proposed in [11] to realize source adaptivity in ICA. With its non-parametric nature the approach offers a valuable alternative to methods which only can realize a limited amount of adaptivity. However we have shown that the QDE can be viewed as a natural extension of such parametric approaches and in particular the flexible ICA method [4] (and similarly the extended infomax algorithm [5]) correspond to special realizations of QDE with lowest complexity. These low-complexity variants are necessary for initial learning in order to stabilize the overall optimization. Finally, the simulation results indicate that already for data sets of moderate size it is possible to use QDE in order to improve the performance in blind source separation as compared with flexible ICA.

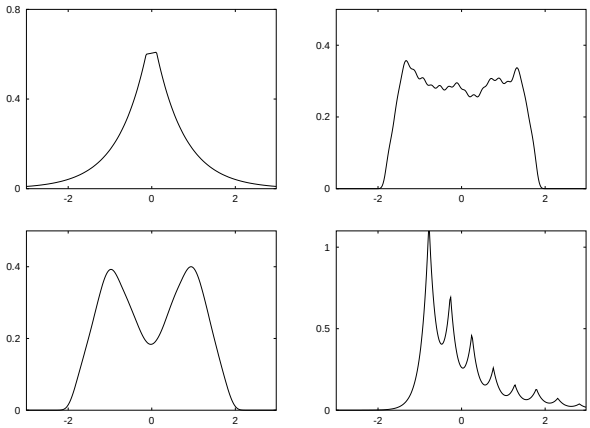

Fig. 2. QDE with $M = 32$ quantization levels, ordered to match the original sources in figure 1

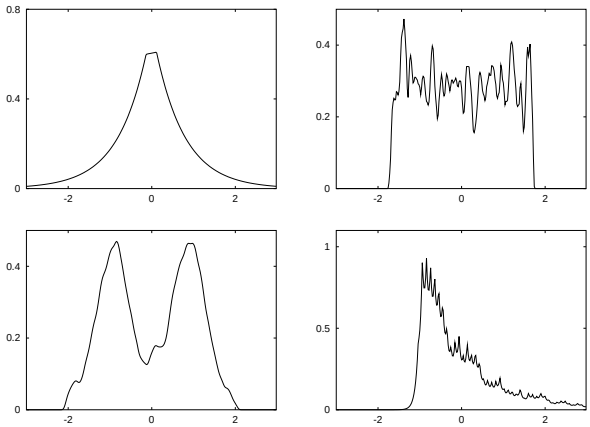

Fig. 3. QDE with $M = 128$ quantization levels

6. REFERENCES

[1] B. Pearlmutter and L. Parra, "A context sensitive generalisation of ica," in *International Conference on Neural Information Processing*, Hong Kong, 1996.

[2] D. Pham and P. Garat, "Blind separation of mixtures of independent sources through a quasi-maximum likelihood approach," *IEEE Trans. on Signal Processing*, , no. 7, pp. 1712–1725, 1997.

[3] J. F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 9, no. 10, pp. 2009–2025, 1998.

[4] S. Choi, A. Cichocki, and S. Amari, "Flexible independent component analysis," in *Neural Networks for Signal Processing VIII*, 1998, pp. 83–92.

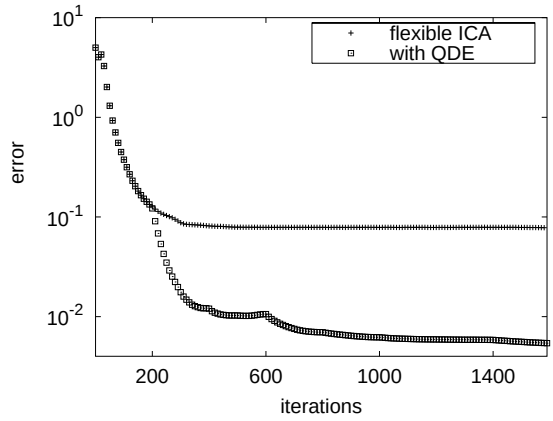

Fig. 4. Separation error of flexible ICA and QDE based ICA

[5] Te-Won Lee, Mark Girolami, and Terrence J. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources," *Neural Computation*, vol. 11, no. 2, pp. 417–441, 1999.

[6] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999, http://www.icsi.berkeley.edu/ jagota/NCS.

[7] S. Amari and J. Cardoso, "Blind source separation – semiparametric statistical approach," *IEEE Trans. Signal Processing*, vol. 45, no. 11, pp. 2692–2700, 1997.

[8] N. Vlassis and Y. Motomura, "Efficient source adaptivity in independent component analysis," *IEEE Transactions on Neural Networks*, vol. 12, no. 3, pp. 559–566, 2001.

[9] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London and New York, 1986.

[10] D. W. Scott, *Multivariate Density Estimation*, Wiley, 1992.

[11] Peter Meinicke and Helge Ritter, "Quantizing density estimators," 2001, submitted to Neural Information Processing Systems.

[12] B. W. Silverman, "Kernel density estimation using the fast Fourier transform," *Applied Statistics*, vol. 31, no. 1, pp. 93–99, 1982.

[13] Stuart Geman and Chii-Ruey Hwang, "Nonparametric maximum likelihood estimation by the method of sieves," *The Annals of Statistics*, vol. 10, no. 2, pp. 401–414, 1982.

[14] Shun ichi Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[15] J. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. on Signal Processing*, , no. 12, pp. 3017–3030, 1996.

[16] Peter Meinicke, *Unsupervised Learning in a Generalized Regression Framework*, Ph.D. thesis, Universitaet Bielefeld, 2000, `http://archiv.ub.uni-bielefeld.de/disshabi/2000/0033/`.