Hybrid Bernstein Normalizing Flows

for Flexible Multivariate Density Regression with Interpretable Marginals

Marcel Arpogaus[1,2] Thomas Kneib[2] Thomas Nagler[3,4] David Rügamer[3,4]

1 HTWG Hochschule Konstanz Technik, Wirtschaft und Gestaltung

2 GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

3 LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

4 MCML Munich Center for Machine Learning

Intro

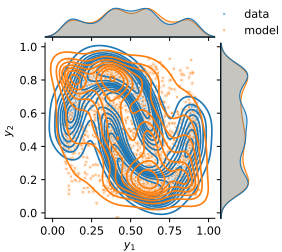


Figure: MCTM (orange) applied to a complex data distribution (blue).

- Multivariate conditional density estimation is a challenging task.
- Neural network based approaches offer great flexibility but lack interpretability.
- Statistical methods focusing on interpretation of feature effects, lack flexibility.

$\Rightarrow$ Let's combine the best of both worlds!

Our Contributions

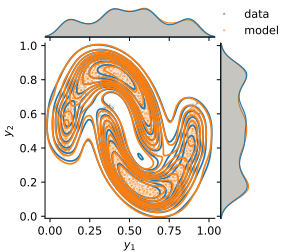
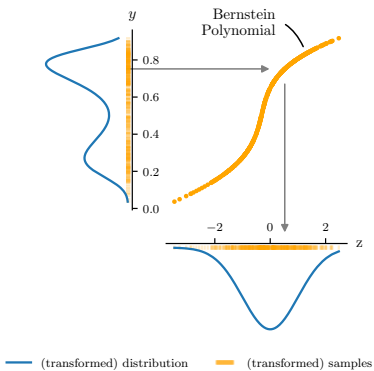


Figure: Our Approach (orange) applied to a the same complex data distribution (blue).

We propose a hybrid approach combining the transparency of multivariate conditional transformation models (MCTMs) [1] with the flexibility of normalizing flows (NFs) [2].

- Understand the impact of features on each response variable within the marginal distribution.
- Simultaneously enable complex modeling of the dependency structure among outcome dimensions.
- We assessed its effectiveness on both simulated and real-world datasets.

Transformation Models



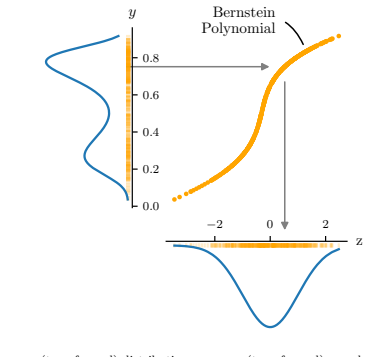
Conditional Transformation Models [3, 4]

- Use flexible, strictly monotone, covariate-dependent transformations $h(y|\mathbf{x})$ to map the data to a reference distribution F_z.
- Allow conditional density estimation under weak assumptions.
- Likelihood is given by the transformation theorem for densities [5]:

$$p_y(y|\mathbf{x}) = p_z\left(h(y|\mathbf{x})\right) \left|\det\left(\nabla h(y|\mathbf{x})\right)\right|$$

- The absolute value of the Jacobian determinant ensures that the probability mass is preserved.

Transformation Models



Bernstein Polynomials

1. Can approximate any continuous function, to any desired accuracy, over a prescribed interval [6].
2. Ability to increase the flexibility at no cost to the training stability [7].
3. Easy to enforce monotonicity [4].
4. Gives smooth approximations even for high order polynomials [6].

Multivariate Conditional Transformation Models [1]

- Lower triangular $(J \times J)$ coefficient matrix for linear dependencies:

$$\Lambda(\mathbf{x}) = \begin{pmatrix} 1 & & & 0 \\ \lambda_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ \lambda_{J1} & \lambda_{J2} & \cdots & 1 \end{pmatrix}$$

- Multiplied with marginal transformations $\tilde{h}_j\left(y_j|\mathbf{x}\right), j = 1, \ldots, J$:

$$h(y|\mathbf{x}) = \Lambda(\mathbf{x}) \left(\tilde{h}_1(y_1|\mathbf{x}), \ldots, \tilde{h}_J(y_J|\mathbf{x})\right)^T$$

- Same structure as a Gaussian copula with parametrization $\Sigma = \Lambda^{-1}\Lambda^T$:

$$\mathbb{P}(\mathbf{Y} \leq \mathbf{y}) = \Phi_{0,\Sigma}\left(\tilde{h}_1(y_1|\mathbf{x}), \ldots, \tilde{h}_J(y_J|\mathbf{x})\right)$$

Interpretation [8, 9, 10]

Shifted Bernstein polynomial

$$h(y|x) = \underbrace{\boldsymbol{\alpha}(y)^\top \boldsymbol{\theta}}_{\text{Bernstein Polynomial}} + x\beta$$

Common interpretational scales depending on the distribution F_Z [8]

F_Z	F_Z^{-1}	Interpretation of shift terms β
Logistic	logit	log odds-ratio
Gompertz	cloglog	log hazard-ratio
Gumbel	loglog	log hazard-ratio for $Y_r = K + 1 - Y$
Normal	probit	not interpretable directly

Masked Autoregressive Flows (MAF) [13]

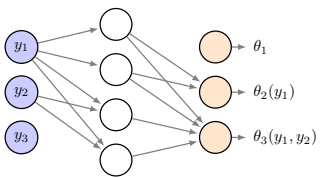


Figure: Masked Autoregressive Neural Network [11]

Factorize the multivariate distributions, based on the chain rule of probability

$$p_y(\mathbf{y}) = \prod_{i=1}^{D} p_z\left(h_i(y_i, \theta_i(\mathbf{y}_{<i}))\right) \left|\det \nabla h_i(y_i, \theta_i(\mathbf{y}_{<i}))\right|$$

The autoregressive property $\mathbf{y}_{<i}$ is enforced through parameter masking in the neural network estimating the parameters θ_i [12].

1. Motivation

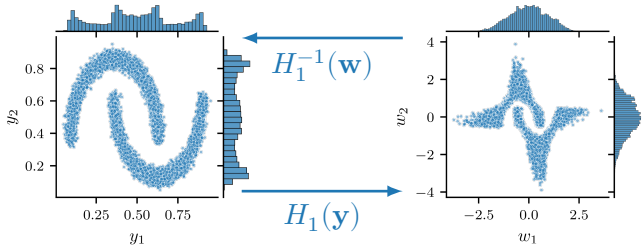
2. Background and Related Works

3. **Proposed Hybrid Models**

4. Experiments

5. Summary

Model Specification

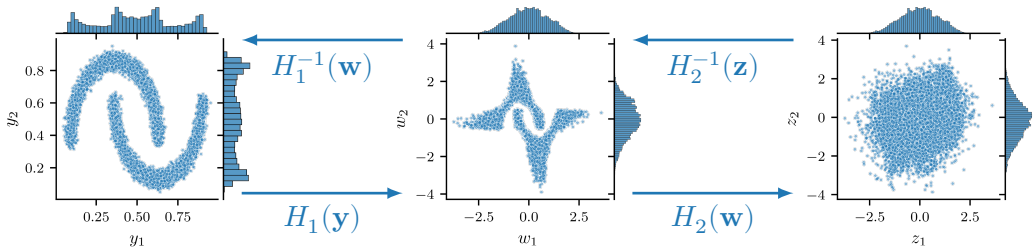


First, we model the conditional marginal distributions $F_{Y_j}(y_j|\mathbf{x})$ using a transformation model:

$$H_1(\mathbf{y}, \mathbf{\Theta}(\mathbf{x})) = \left(h_1(y_1, \boldsymbol{\theta}_{1,\mathbf{x}}), \dots, h_1(y_J, \boldsymbol{\theta}_{J,\mathbf{x}})\right)^\top = (w_1, \dots, w_J)^\top$$

The parameters for this transformation can be modeled in the same interpretable fashion as the original MCTM allow.

Model Specification



We model the dependencies between elements of $\mathbf{W}$ using an autoregressive flow:

$$H_2(\mathbf{w}, \mathbf{\Psi}(\mathbf{w}, \mathbf{x})) = \Big(w_1, h_2(w_2|\boldsymbol{\psi}_{2,w_1,\mathbf{x}}), \dots, h_2(w_J|\boldsymbol{\psi}_{J,\mathbf{w}_{<J},\mathbf{x}})\Big) = (z_1, \dots, z_J)^\top$$

The parameters $\boldsymbol{\psi}_{j,\mathbf{w}_{<j},\mathbf{x}}$ of the transformation functions $h_2(\cdot)$ are estimated by a masked neural network depending on previous elements of $\mathbf{w}$ and covariates $\mathbf{x}$.

Parametrization and Interpretability

Marginal Transformation (H_1)

- Can utilizes shifted Bernstein polynomials: $h(y|x) = \boldsymbol{\alpha}(y)^\top \boldsymbol{\theta} + x\beta$
- Same interpretational scale as MCTMs.
- Example: With a logistic base distribution, linear effect coefficients represent changes in log-odds ratios.

Autoregressive Flow (H_2)

- Parameters are estimated by a masked neural network.
- Models complex dependencies but lacks direct interpretability.

$\Rightarrow$ We prioritizes marginal interpretability.

Model Training and Inference

Optimize the model parameters $\boldsymbol{\omega}$

- Minimize the conditional negative log-likelihood:

$$\text{NLL}(\boldsymbol{\omega}|\mathcal{D}) = -\sum_{(\mathbf{y},\mathbf{x})\in\mathcal{D}} \log f_Z\left(h_{\boldsymbol{\omega}}\left(\mathbf{y}|\mathbf{x}\right)\right)\left|\nabla h_{\boldsymbol{\omega}}\left(\mathbf{y}|\mathbf{x}\right)\right|$$

- H_1 and H_2 are trained separately for optimal results

Sample from the Learned Distribution

1. Sample $\mathbf{z}$ from the base distribution F_Z.
2. Apply the inverse autoregressive flow: $\mathbf{w} = H_2^{-1}(\mathbf{z}|\mathbf{x})$.
3. Apply the inverse marginal transformation: $\mathbf{y} = H_1^{-1}(\mathbf{w}|\mathbf{x})$.

In short: $\mathbf{y} = H_1^{-1}(H_2^{-1}(\mathbf{z}|\mathbf{x})|\mathbf{x})$ with $\mathbf{z} \sim F_Z$.

1. Motivation

2. Background and Related Works

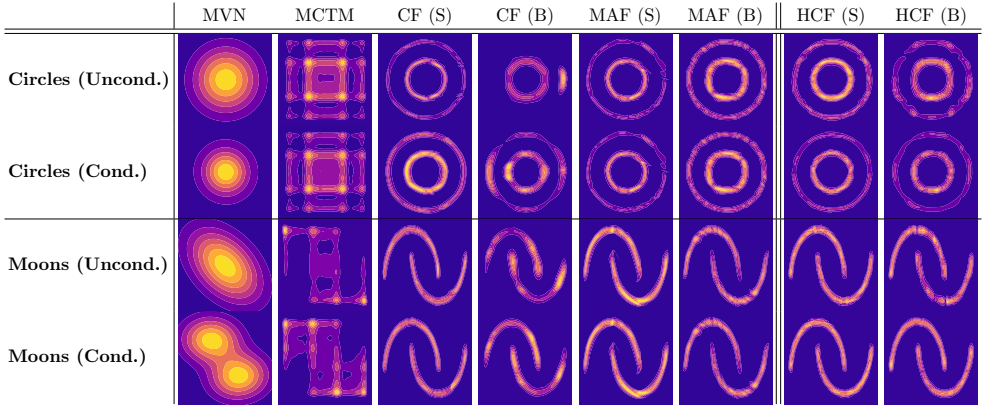
3. Proposed Hybrid Models

4. **Experiments**

5. Summary

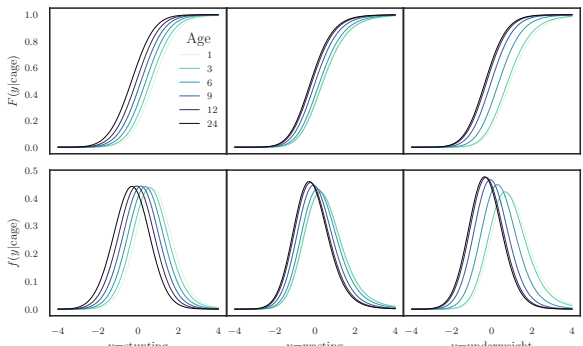
Simulated Data

Comparing: Multivariate Normals (MVN), Multivariate Conditional Transformation Models (MCTM), Coupling Flows (CF), MAF, and Hybrid Coupling Flows (HCF). HCF combines element-wise Bernstein polynomials for marginals with a coupling layer for dependencies using either Bernstein Polynomials (B) or quadratic Splines (S).



Malnutrition Data

Child malnutrition data from India, modeling the joint distribution of three indicators (`stunting`, `wasting`, `underweight`) conditional on child age (`cage`).



Marginal effects indicate nonlinear shifts toward lower nutrition status with increasing age.

Malnutrition Data

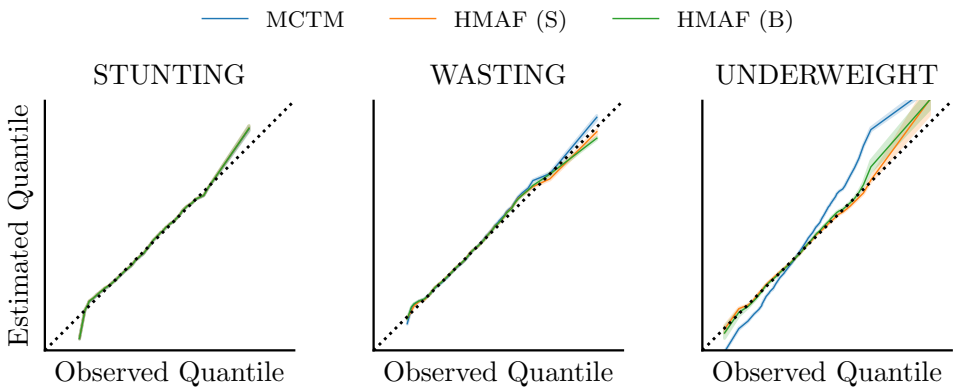


Figure: QQ plot comparing empirical quantiles of the dataset with those generated by the three models. The solid lines represent the mean, while the shaded areas indicate the 95% probability intervals obtained from 20 trials of randomly initialized models.

1. Motivation
2. Background and Related Works
3. Proposed Hybrid Models
4. Experiments
5. **Summary**

Summary

- Our hybrid approach combines the interpretability of CTMs with the flexibility of autoregressive NFs.
- results on simulated and real-world datasets showed that our method is competitive with state-of-the-art methods.
- Ideal for understanding individual feature effects while modeling complex response variable relationships.

Open Questions

- Understanding feature effects on the dependence structure in H_2 remains an open research question.
- Possible analyses include examining flow parameter variations with feature values or employing xAI techniques for insight.

Thank you!

Do you have any Questions?

Feel free to get in touch!

Poster Session 1 (Today, 16:00-18:00)

LinkedIn /in/MArpogaus

GitHub /MArpogaus

Mail marcel.arpogaus@htwg-konstanz.de

Source Code `https://github.com/MArpogaus/hybrid-flows`

Normalizing Flows [12, 2]

In the Machine Learning literature, transformation models are defined as a composition $h^{-1}(z) = f_K \circ f_{K-1} \circ ... \circ f_1(y)$ of K simple transformation functions f_i, to model a generative process, known as Normalizing Flow [2].

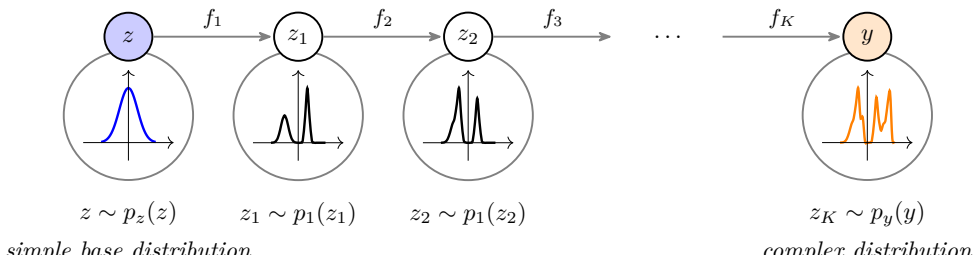
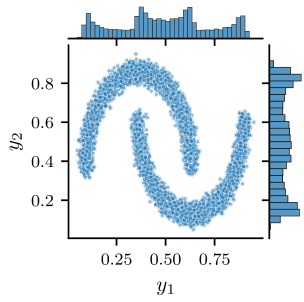
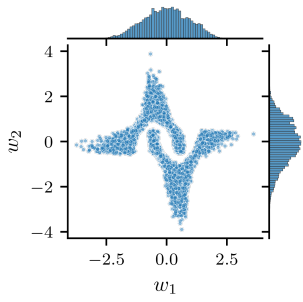


Figure: Illustration of a normalizing flow, transforming a simple distribution $p_z(z)$ to a more complex one. Illustration inspired by [14]

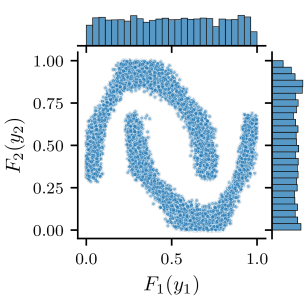
Relation to Copula Methods



(a) Original Data



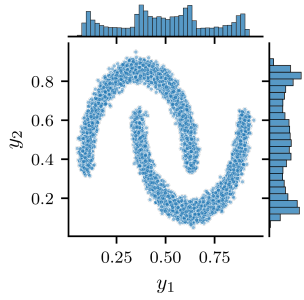
(b) Normalized Marginals



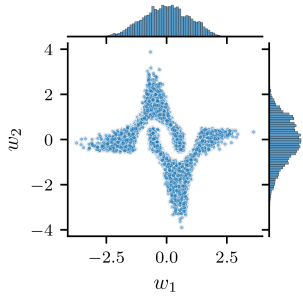
(c) Uniform Marginals

- The first step (H_1) models marginals $F_{Y_j|\mathbf{X}}$ and transforms them to the base distribution F_Z.
- Applying the PIT, $u_j = F_Z(z_{1j})$, yields uniform marginals.

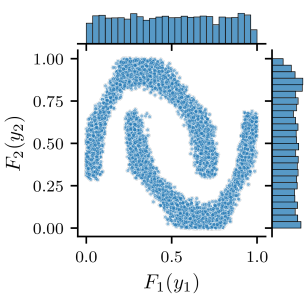
Relation to Copula Methods



(a) Original Data



(b) Normalized Marginals

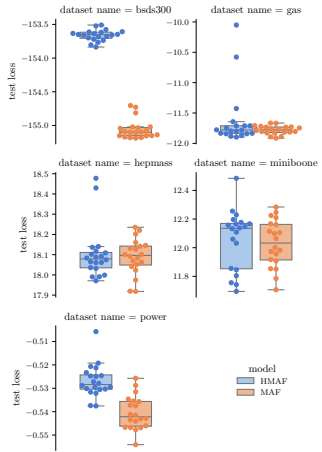


(c) Uniform Marginals

The conditional copula density c of $\mathbf{u} = (u_1, \ldots, u_J)^\top$ is given by:

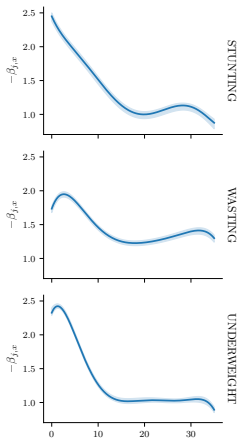
$$c(\mathbf{u}|\mathbf{x}) = \frac{f_{\mathbf{Y}|\mathbf{X}}(F^{-1}_{Y_1|\mathbf{X}}(u_1|\mathbf{x}), \ldots, F^{-1}_{Y_J|\mathbf{X}}(u_J|\mathbf{x})|\mathbf{x})}{\prod_{j=1}^{J} f_{Y_j|\mathbf{X}}(F^{-1}_{Y_j|\mathbf{X}}(u_j|\mathbf{x})|\mathbf{x})}.$$

Benchmark datasets



- We evaluate our method on five benchmark datasets (POWER, GAS, HEPMASS, MINIBOONE and BSDS300) and compare *Masked Autoregressive Flows* (*MAF*) with our *Hybrid Masked Autoregressive Flows* (*HMAF*).
- *HMAF* generally provide comparable results compared to *MAF*.

Malnutrition: Inverse Marginal Shift



- Inverse marginal shift terms reveal complex nonlinear age effects, more pronounced for stunting and underweight.
- Acute malnutrition (as measured by the stunting indicator) materializes more quickly than chronic malnutrition (as measured by the wasting indicator).
- Underweight represents a mixture of both acute and chronic malnutrition, which again fits with the estimated shift term.

Malnutrition Marginal Transformation

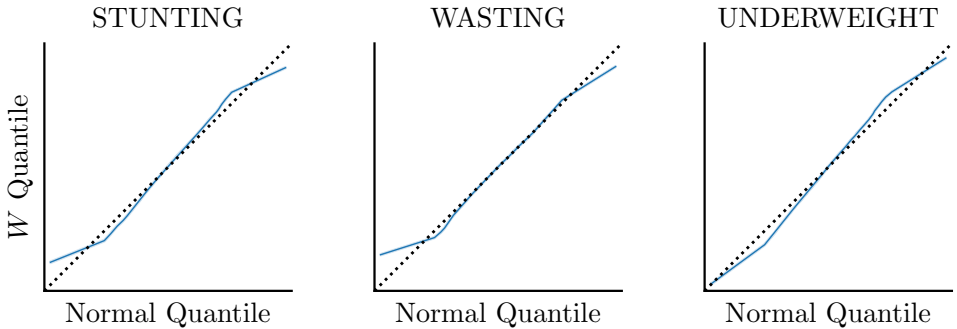


Figure: QQ plots of transformed samples against a standard normal distribution. Deviations from the diagonal indicate non-normality. The solid line represents the mean, while the shaded area indicates the 95% probability intervals obtained from 20 trials of randomly initialized models.

Runtime of Model Variants

Runtime for training and evaluating our models on the HPC Cluster at the University of Applied Sciences Esslingen, utilizing NVIDIA L40S GPUs with 48 GB VRAM.

Table: Runtime in Minutes for training and evaluation of models on benchmark data. Variance resulting deviations from 20 runs reported as standard deviation.

model	dataset name	train	evaluation
HMAF	bsds300	1191.992 ± 537.944	481.991 ± 0.650
	gas	319.809 ± 132.472	14.931 ± 0.043
	hepmass	229.736 ± 155.484	15.094 ± 0.047
	miniboone	82.882 ± 58.692	3.933 ± 0.012
	power	437.108 ± 63.707	11.321 ± 0.091
MAF	bsds300	261.977 ± 70.957	16.716 ± 0.022
	gas	68.993 ± 0.073	1.858 ± 0.005
	hepmass	34.774 ± 0.003	1.540 ± 0.004
	miniboone	16.486 ± 1.404	0.279 ± 0.001
	power	136.796 ± 0.004	4.979 ± 0.120

Runtime of Model Variants

Runtime for training and evaluating our models on the HPC Cluster at the University of Applied Sciences Esslingen, utilizing NVIDIA L40S GPUs with 48 GB VRAM.

Table: Mean runtime in seconds for training and evaluation of models on malnutrition data. Variance resulting deviations from 20 runs reported as standard deviation.

model	training	evaluation
HMAF (B)	260.752 ± 121.895	20.823 ± 0.535
HMAF (S)	1993.317 ± 717.933	19.649 ± 0.110
MCTM	4106.187 ± 725.136	16.877 ± 0.847

[1] N. Klein, T. Hothorn, L. Barbanti, and T. Kneib, "Multivariate conditional transformation models," *Scandinavian Journal of Statistics*, vol. 49, no. 1, pp. 116–142, 2022, ISSN: 1467-9469. DOI: 10.1111/sjos.12501.

[2] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing Flows for Probabilistic Modeling and Inference," *Journal of Machine Learning Research*, vol. 22, no. 57, pp. 1–64, 2021, ISSN: 1533-7928. [Online]. Available: http://jmlr.org/papers/v22/19-1028.html.

[3] T. Hothorn, T. Kneib, and P. Bühlmann, "Conditional Transformation Models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 76, no. 1, pp. 3–27, Jan. 2014, ISSN: 1369-7412. DOI: 10.1111/rssb.12017.

[4] T. Hothorn, L. Möst, and P. Bühlmann, "Most Likely Transformations," *Scandinavian Journal of Statistics*, vol. 45, no. 1, pp. 110–134, 2018, ISSN: 1467-9469. DOI: 10.1111/sjos.12291.

[5] T. Kneib, A. Silbersdorff, and B. Säfken, "Rage Against the Mean – A Review of Distributional Regression Approaches," *Econometrics and Statistics*, Aug. 10, 2021, ISSN: 2452-3062. DOI: 10.1016/j.ecosta.2021.07.006.

[6] R. T. Farouki, "The Bernstein Polynomial Basis: A Centennial Retrospective," *Comput. Aided Geom. Des.*, vol. 29, no. 6, pp. 379–419, Aug. 2012, ISSN: 0167-8396. DOI: 10.1016/j.cagd.2012.03.001.

[7] S. Ramasinghe, K. Fernando, S. Khan, and N. Barnes, "Robust normalizing flows using Bernstein-type polynomials," arXiv: 2102.03509 [cs, stat], pre-published.

[8] L. Kook, L. Herzog, T. Hothorn, O. Dürr, and B. Sick, "Deep and interpretable regression models for ordinal outcomes," *Pattern Recognition*, vol. 122, p. 108 263, Feb. 1, 2022, ISSN: 0031-3203. DOI: 10.1016/j.patcog.2021.108263.

[9] G. Tutz, *Regression for Categorical Data* (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge: Cambridge University Press, 2011, ISBN: 978-1-107-00965-3. DOI: 10.1017/CBO9780511842061.

[10] B. Sick and O. Dürr, "Interpretable Neural Causal Models with TRAM-DAGs," arXiv: 2503.16206 [stat], pre-published.

[11] M. Germain, K. Gregor, I. Murray, and H. Larochelle, "MADE: Masked Autoencoder for Distribution Estimation," in *Proceedings of the 32nd International*

Conference on Machine Learning, PMLR, Jun. 1, 2015, pp. 881–889. [Online]. Available: https://proceedings.mlr.press/v37/germain15.html.

[12] I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing Flows: An Introduction and Review of Current Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3964–3979, Nov. 2021, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2020.2992934. arXiv: 1908.09257.

[13] G. Papamakarios, T. Pavlakou, and I. Murray, "Masked Autoregressive Flow for Density Estimation," Jun. 14, 2018. arXiv: 1705.07057 [cs, stat].

[14] L. Weng, "Flow-based deep generative models," lilianweng.github.io/lil-log. [Online]. Available: http://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html.