A Stochastic Variational Inference Approach for Semiparametric Distributional Regression

Final Kolloquium MSc Applied Statistic

Sebastian Lorek

First supervisor: Prof. Dr. Thomas Kneib
Second supervisor: Gianmarco Callegher

2023-11-28

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Table of Contents I

Introduction

Bayesian Inference

- Focal point of interest is the posterior distribution
- General posterior specification for Bayesian regression

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{y}, \mathcal{D}) &= \frac{p(\mathbf{y}, \boldsymbol{\theta}|\mathcal{D})}{p(\mathbf{y}|\mathcal{D})} \\
&= \frac{p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{D})p(\boldsymbol{\theta})}{p(\mathbf{y}|\mathcal{D})} \\
&= \frac{p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{D})p(\boldsymbol{\theta})}{\int p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{D})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \\
&\propto p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{D})p(\boldsymbol{\theta})
\end{aligned}
$$

- Primary challenge is the calculation of the normalizing constant/evidence

- MCMC methods have been **so far** the most common methods for Bayesian inference
  - Enjoy nice properties
  - Unfortunately do not scale well for modern applications
- Trade some **accuracy** for **scalability**
- Make use of SGD and automatic differentiation

Theory

Variational Inference

- VI is a method from machine learning to **approximate** probability densities (Jordan et al. 1999)
- Approximate posterior with a variational distribution $q(\boldsymbol{\theta}|\phi)$ from a predefined parametric variational family $\mathcal{Q}$

$$q(\boldsymbol{\theta}|\phi) \in \mathcal{Q}$$

- Use optimization (SGD) to find a member that is as closely as possible to the true posterior
- What means close in terms of distributions ?

Kullback-Leibler divergence

- A well known divergence measure is the Kullback-Leibler divergence (Kullback and Leibler 1951)
- Quantifies the proximity between two probability distributions

$$\begin{aligned} \mathrm{D_{KL}}\left(q(\boldsymbol{\theta}|\phi)||p(\boldsymbol{\theta}|\mathbf{y},\mathcal{D})\right) &= \int q(\boldsymbol{\theta}|\phi) \ln\left(\frac{q(\boldsymbol{\theta}|\phi)}{p(\boldsymbol{\theta}|\mathbf{y},\mathcal{D})}\right) d\boldsymbol{\theta} \\ &= \mathrm{E}_{q(\boldsymbol{\theta}|\phi)}\left[\ln\left(\frac{q(\boldsymbol{\theta}|\phi)}{p(\boldsymbol{\theta}|\mathbf{y},\mathcal{D})}\right)\right] \end{aligned}$$

- In short $\mathrm{D_{KL}}\left(q||p\right)$
- It holds that $\mathrm{D_{KL}}\left(q||p\right) \geq 0$
- Has some nice properties but also drawbacks (not a distance)

Optimization objective

$$\hat{\phi} = \arg\min_{\phi} \mathrm{E}_{q(\boldsymbol{\theta}|\phi)} \left[\ln \left(\frac{q(\boldsymbol{\theta}|\phi)}{p(\boldsymbol{\theta}|\mathbf{y}, \mathcal{D})} \right) \right]$$

- Make q as close as possible to p

- The posterior is unknown, but the **evidence** is a **constant** in the optimization and thus cancels out
- Allows us to rewrite the objective as

$$
\begin{aligned}
\hat{\phi} &= \arg\max_{\boldsymbol{\phi}} \mathrm{E}_{q(\boldsymbol{\theta}|\phi)}\left[\ln\left(\frac{p(\mathbf{y}, \boldsymbol{\theta}|\mathcal{D})}{q(\boldsymbol{\theta}|\phi)}\right)\right] \\
&= \arg\max_{\boldsymbol{\phi}} \mathrm{ELBO}(\phi)
\end{aligned}
$$

Evidence Lower Bound

- What does the ELBO ?

$$\mathrm{ELBO}(\phi) = \mathsf{E}_{q(\boldsymbol{\theta}|\phi)}\left[\ln(p(\mathbf{y}|\mathcal{D}, \boldsymbol{\theta}))\right] - \mathsf{D}_{\mathsf{KL}}\left(q(\boldsymbol{\theta}|\phi)||p(\boldsymbol{\theta})\right)$$

- Actually something similar to MAP/ML 🤔

Variational family

- What is the $\mathcal{Q}$ and thus $q(\boldsymbol{\theta}|\phi)$?
- Flexibility of $\mathcal{Q}$ and thus $q(\boldsymbol{\theta}|\phi)$ significantly influences the optimization
  - Complex $\mathcal{Q} \rightarrow$ better approximations, but increased complexity during optimization
  - Simple $\mathcal{Q} \rightarrow$ worse approximations, but allows for easier optimization

- We use structured mean-field variational inference (Wainwright and Jordan 2007)

$$q(\boldsymbol{\theta}|\phi) = \prod_{j=1}^{J} q_j(\boldsymbol{\theta}_j|\phi_j)$$

- One q_j (factor) for each parameter block
- Allows for interdependencies in parameter blocks

- How are the q_j defined ?

$$\boldsymbol{\theta}_j = \mathbf{L}_j \boldsymbol{\epsilon}_j + \boldsymbol{\mu}_j, \ \boldsymbol{\epsilon}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$\boldsymbol{\theta}_j \sim \mathcal{N}(\boldsymbol{\mu}_j, \mathbf{L}_j \mathbf{L}_j^{\mathrm{T}})$$

- Called "reparameterization trick" $\boldsymbol{\theta}_j = \mathbf{g}_{\phi_j}(\boldsymbol{\epsilon}_j)$
- For positive restricted parameters we need to chain another transformation layer via exp transformation

Stochastic variational inference

- SGD to optimize the ELBO (Hoffman et al. 2012)
- Only use a subset $\mathcal{I}$ of the data in each iteration (ELBO remains unbiased)

“Black-box” variational inference

- Using SVI with the “reparameterization trick”
- Evaluate the integral in the ELBO with Monte Carlo integration
- Calculate $\nabla_\phi \text{ELBO}(\phi)_\mathcal{I}$ with automatic differentiation
- Researcher only formulates a probabilistic model and provides a dataset (Kucukelbir et al. 2016), inference algo. is model “agnostic”

Optimization

- Split data into training and validation
- Decide for batch-size (subset of the data) and variational sample size (Monte Carlo integration)
- Optimize the ELBO using training data and monitor convergence in the validation dataset

$$\hat{\phi}^t = \hat{\phi}^{t-1} + \rho_t \nabla_\phi \text{ELBO}(\phi)_{\mathcal{I}^t}\Big|_{\phi=\hat{\phi}^{t-1}}$$

- We use Adam (Kingma and Ba 2014) as our optimizer

Impact Analysis

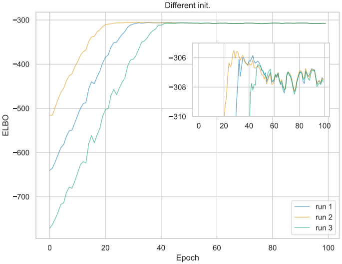

Figure 1: ELBO traces for 3 different SGD runs, using different initializations but the same seed. We use a batch size of 128, 64 samples from the variational distribution and a learning rate of 1e-2.

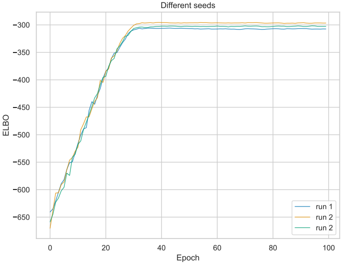

Figure 2: ELBO traces for 3 different SGD runs, using different seeds but the same initialization. Otherwise same configuration as in Figure 1.

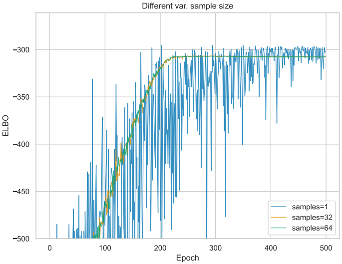

Figure 3: ELBO traces for 3 different SGD runs, using different variational sample sizes. We use batch VI with a learning rate of 1e-2 and the same seed as in Figure 1.

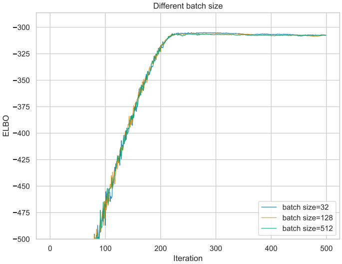

Figure 4: ELBO traces for 3 different SGD runs, using different batch sizes. We use a variational sample size of 64 with a learning rate of 1e-2 and the same seed as in Figure 1.

Semiparametric distributional regression

$$y_i \overset{\text{ind.}}{\sim} p(\vartheta_{i,1}, \dots, \vartheta_{i,g})$$
$$\vartheta_{i,l} = h_l^{-1}(\eta_{i,l}),\ l = 1, \dots, g$$

- Response distributions beyond the normal distribution
- Linear predictors with inverse link functions
- Structured addtive linear predictors so fixed and smooth effects
    - B-spline basis functions
- Augmented with priors
    - Bayesian P-splines

Application

Implementation

- Developed a small `python` package `tigerpy`, which consists of two libraries
- A model building library `tigerpy.model`
  - Construct the model
  - Uses the idea of probabilistic graphical models
- An inference library `tigerpy.bbvi`
  - Runs the BBVI inference algorithm
- Aligns with concepts found in `liesel` (Riebl et al. 2022)

Simulation Studies

- Conducted two simulation studies
- First studies asymptotic behavior of the posterior means of BBVI
- Second targets the posterior distributions as a whole and compares BBVI with MCMC

Consistency study

- Study the asymptotic behavior of the posterior means
- Performance measures include bias and empirical standard error
- Simulation repetitions: $n_{\text{sim}} = 200$
- Two models
    1. M_1: Bayesian linear regression
    2. M_2: Bayesian logistic regression

- Two data generating processes, $n_{\text{obs}} = 50, 100, 500, 1000, 5000$
- For M_1

$$\begin{aligned} y_i | x_i &\sim \mathcal{N}(3.0 + 0.2x_i - 0.5x_i^2, 1.0^2), \\ x_i &\sim \mathcal{U}(-3, 3),\ i = 1, \dots, n_{\text{obs}} \end{aligned}$$

- For M_2

$$\begin{aligned} y_i | x_i &\sim \text{Bern}(\sigma(1.0 + 2.0x_i)), \\ x_i &\sim \mathcal{U}(-3, 3),\ i = 1, \dots, n_{\text{obs}} \end{aligned}$$

M_1

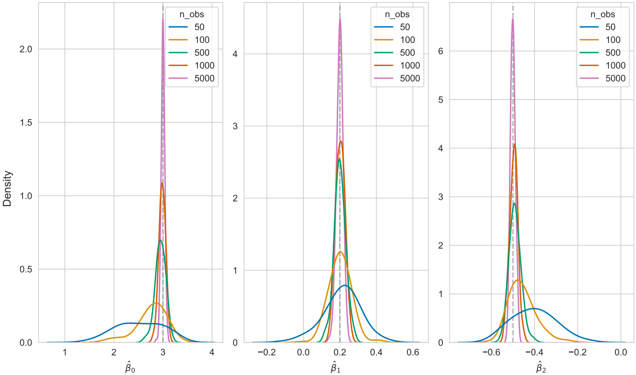

Figure 5: Kernel density for the posterior means of the location parameters of M_1, true parameters given by $\boldsymbol{\beta} = [3.0, 0.2, -0.5]^{\mathrm{T}}$ (grey dashed line).

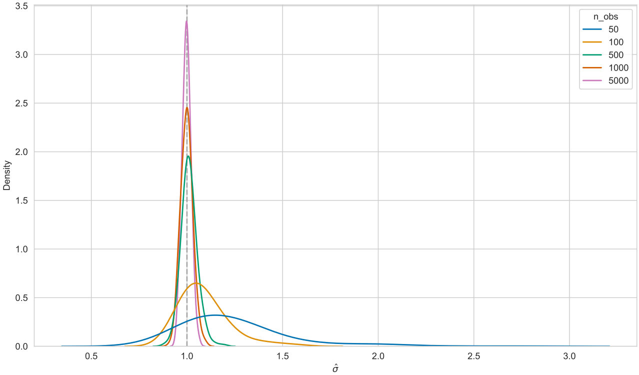

Figure 6: Kernel density for the posterior means of the scale parameter of M_1, true parameter given by $\sigma = 1.0$ (grey dashed line).

Table 1: Simulation results for the parameters of M_1.

	Bias					EmpSE				
n_{obs}	50	100	500	1000	5000	50	100	500	1000	5000
β_0	**-0.4819**	**-0.2319**	**-0.0520**	**-0.0162**	0.0002	0.4563	0.3046	0.1036	0.0640	0.0344
	(0.0323)	(0.0215)	(0.0073)	(0.0045)	(0.0024)	(0.0229)	(0.0153)	(0.0052)	(0.0032)	(0.0017)
β_1	0.0113	0.0016	0.0003	0.0021	-0.0007	0.0996	0.0598	0.0300	0.0237	0.0166
	(0.0070)	(0.0042)	(0.0021)	(0.0017)	(0.0012)	(0.0050)	(0.0030)	(0.0015)	(0.0012)	(0.0008)
β_2	**0.0972**	**0.0428**	**0.0095**	**0.0033**	-0.0002	0.0975	0.0614	0.0264	0.0176	0.0104
	(0.0069)	(0.0043)	(0.0019)	(0.0012)	(0.0007)	(0.0049)	(0.0031)	(0.0013)	(0.0009)	(0.0005)
σ	**0.2389**	**0.0876**	**0.0115**	-0.0014	-0.0021	0.2851	0.1293	0.0409	0.0300	0.0216
	(0.0202)	(0.0091)	(0.0029)	(0.0021)	(0.0015)	(0.0143)	(0.0065)	(0.0021)	(0.0015)	(0.0011)

Corresponding Monte Carlo SEs are provided below in parentheses; Bias estimates that do not cover 0 in their 95% CI are shown in bold; $n_{sim} = 200$.

M_2

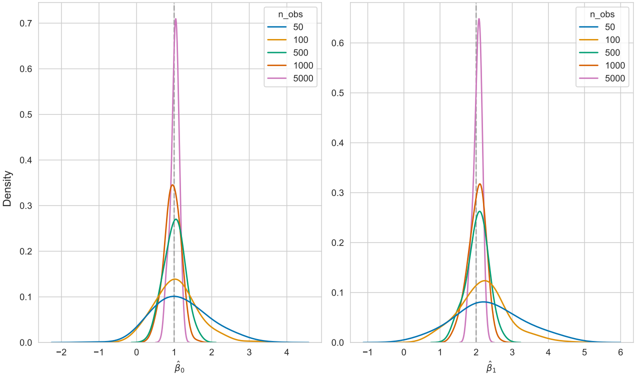

Figure 7: Kernel density for the posterior means of the logit parameters of M_2, true parameters given by $\beta = [1.0, 2.0]^{\mathrm{T}}$ (grey dashed line).

Table 2: Simulation results for the parameters of M_2.

	Bias					EmpSE				
n_{obs}	50	100	500	1000	5000	50	100	500	1000	5000
β_0	**0.1904** (0.0508)	0.0631 (0.0394)	0.0044 (0.0186)	**-0.0481** (0.0147)	0.0137 (0.0078)	0.7180 (0.0360)	0.5568 (0.0279)	0.2632 (0.0132)	0.2072 (0.0104)	0.1103 (0.0055)
β_1	**0.3578** (0.0646)	**0.1914** (0.0462)	**0.0354** (0.0205)	0.0111 (0.0166)	**0.0368** (0.0083)	0.9137 (0.0458)	0.6532 (0.0327)	0.2905 (0.0146)	0.2349 (0.0118)	0.1167 (0.0058)

Corresponding Monte Carlo SEs are provided below in parentheses; Bias estimates that do not cover 0 in their 95% CI are shown in bold; $n_{sim} = 200$.

Posterior density study

- Estimate a smooth function through Bayesian P-splines
- Compare posterior distributions of BBVI (`tigerpy`) and MCMC (`liesel` (ibid.))
- For comparison we use:
  1. Kernel density plots
  2. Wasserstein distance (Kantorovich 1960)
- Generate 4 MCMC Chains and 400 (n_{sim}) BBVI runs
- Data generating process (DGP)

$$
\begin{aligned}
y_i | x_i &\sim \mathcal{N}(f(x_i), 1.5^2) \\
f(x_i) &= 3.0 + 1.75 \sin(1.5 x_i) \\
x_i &\sim \mathcal{U}(-10, 10), \; i = 1, \ldots, 1000,
\end{aligned}
$$

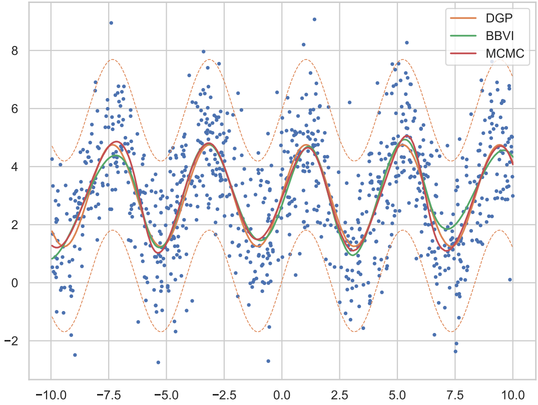

Figure 8: The DGP and the estimated smooth functions from BBVI and MCMC, using the posterior means.

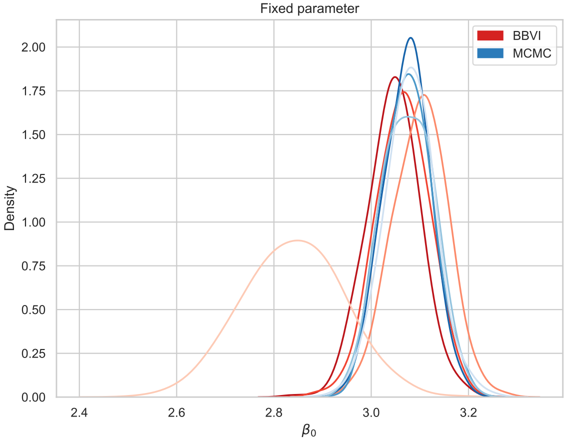

Figure 9: Kernel density for the posterior samples of the fixed intercept β_0, using 4 randomly selected runs from BBVI (red) and 4 chains from MCMC (blue).

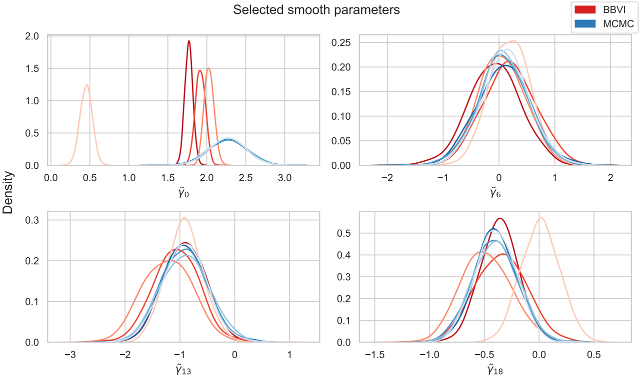

Figure 10: Kernel density for the posterior samples of selected internal spline coefficients $\tilde{\gamma}$, using 4 randomly selected runs from BBVI (red) and 4 chains from MCMC (blue).

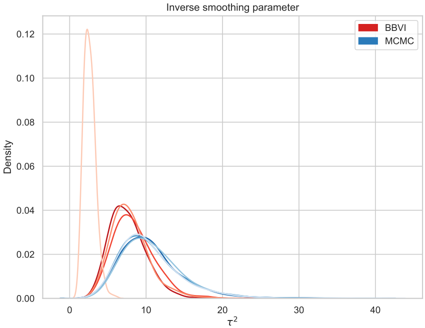

Figure 11: Kernel density for the posterior samples of the inverse smoothing parameter τ^2, using 4 randomly selected runs from BBVI (red) and 4 chains from MCMC (blue).

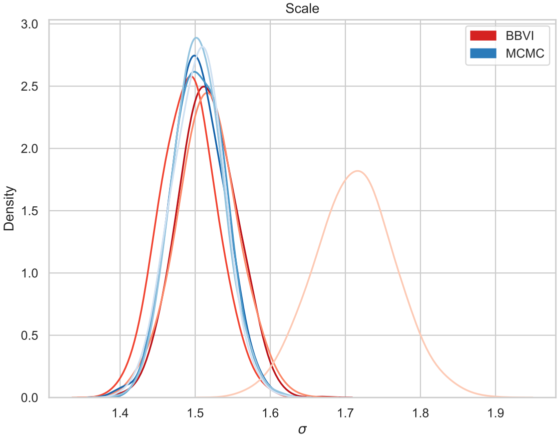

Figure 12: Kernel density for the posterior samples of the scale σ, using 4 randomly slected runs from BBVI (red) and 4 chains from MCMC (blue).

- As a formal measure we use the Wasserstein distance with the "squared euclidean distance" (W_2)
- Allows to compare the "distance" between two probability distributions

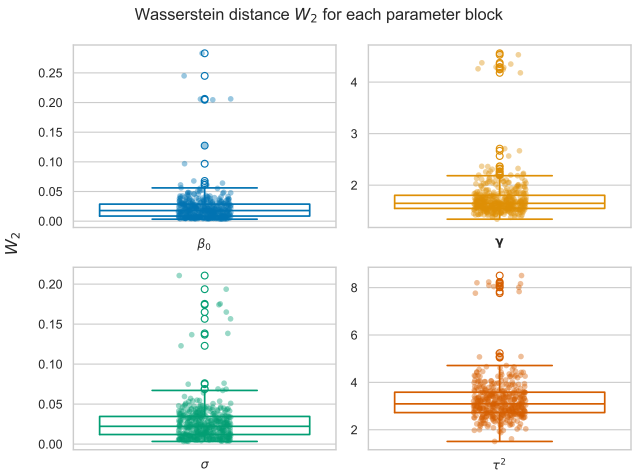

Figure 13: Box plots displaying the Wasserstein distance for the different model parameters.

Open Problems

- Starting with "arbitrary" model initializations for models with scale or shape parameters is numerically too unstable
- Likelihood in the ELBO tends to infinity for "unlikely" samples from the variational distribution
- Forcing the variance of the variational distribution down works ...

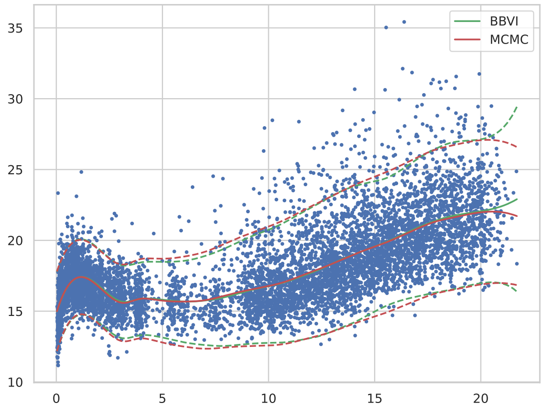

Figure 14: Location scale regression with the Dutch boys dataset, comparing MCMC and BBVI.

- Currently working on a two stage procedure
    1. Start with MAP for a few iterations
        - Caculate Laplace approximation and use it as the initialization for BBVI
    2. Continue with BBVI
- Optional to include one further simulation study or compare results from the first study with MCMC

Hoffman, Matt, David M. Blei, Chong Wang, and John Paisley (2012). *Stochastic Variational Inference*.

Jordan, Michael I., Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul (1999). "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2, pp. 183–233.

Kantorovich, L. V. (July 1960). "Mathematical Methods of Organizing and Planning Production". In: *Management Science* 6.4, pp. 366–422.

Kingma, Diederik P. and Jimmy Ba (2014). *Adam: A Method for Stochastic Optimization*.

Kucukelbir, Alp, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei (2016). *Automatic Differentiation Variational Inference*.

Kullback, S. and R. A. Leibler (Mar. 1951). "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1, pp. 79–86.

Riebl, Hannes, Paul F. V. Wiemann, and Thomas Kneib (2022). *Liesel: A Probabilistic Programming Framework for Developing Semi-Parametric Regression Models and Custom Bayesian Inference Algorithms*.

Wainwright, Martin J. and Michael I. Jordan (2007). "Graphical Models, Exponential Families, and Variational Inference". In: *Foundations and Trends® in Machine Learning* 1.1–2, pp. 1–305.