

Interfacing Bayesian Statistics and Bayes Deep Learning

Jürgen Pilz

Institut für Statistik, Universität Klagenfurt
Universitätsstr. 65-67, 9020 Klagenfurt, Austria
`juergen.pilz@aau.at`

`www.jpilz.net`

World Conference on Data Science & Statistics
June 16-18, 2025 / Amsterdam, The Netherlands

I. Introduction

Over the last years, the number of conferences with joint theme "**Statistics** and Data Science" has been growing.

I am glad that the dinosaur Statistics has not been drowning in the inflation of Machine Learning Conferences.

Good message: Number of research monographs and textbooks on Data Science with a solid treatment of statistics and university curricula of Data Science study programs incl. sound education in statistics are gaining ground!

Essential role of Probability Theory, Information Theory and Statistics is getting more and more acknowledged!

The words of one of the grandfathers of Statistics are still valid:

“Statistics is the grammar of science.”

Karl Pearson (1892)

Introduction

Many advocates of Data Science/Machine Learning forget about/are not aware of the roots of their discipline:

Please thoroughly read

David Donoho (2017): 50 Years of Data Science,

Journal of Computational and Graphical Statistics, 26:4, 745-766,

DOI: 10.1080/10618600.2017.1384734



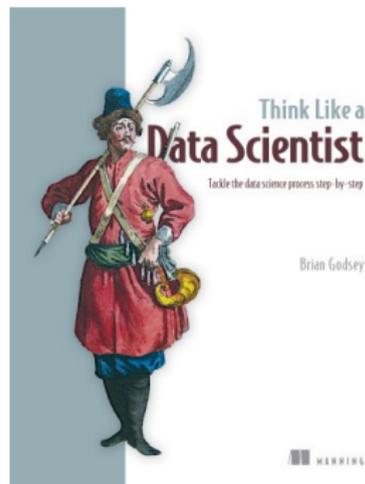
Different Backgrounds

Data Scientists usually come from an engineering background

Statisticians have been trained at Mathematics Departments with specialization in Statistics

Classical (Frequentist) vs. Bayesian Statistics

The Bayesians Have Won Data Science

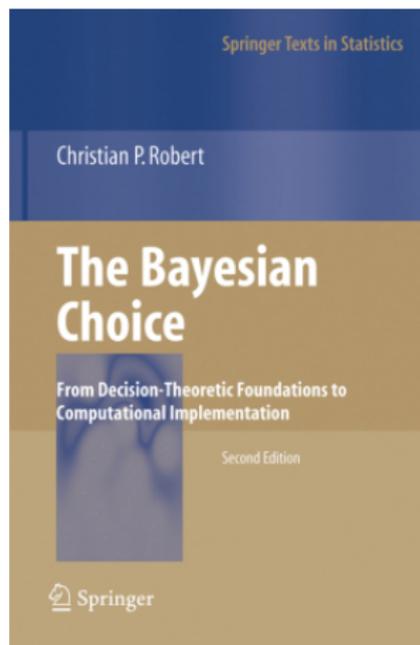


Brian Godsey: Think Like a Data Scientist. Manning Publications 2027

Bayesian Origins

However, it has been a long race!

Good historic account in Christian Robert's (Springer 1997) book



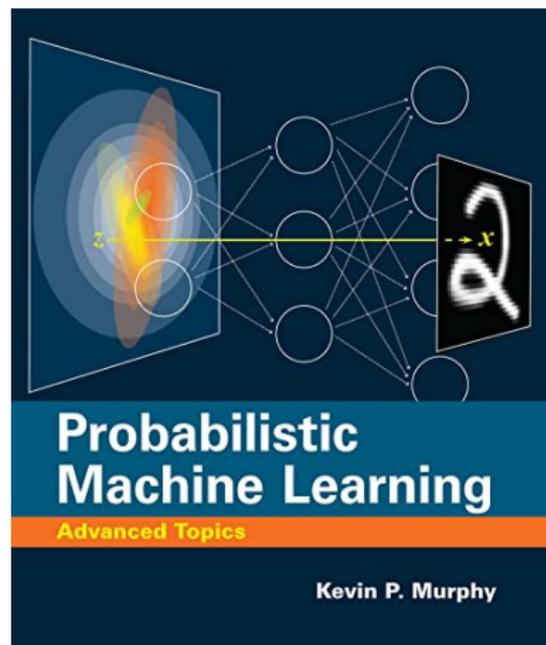
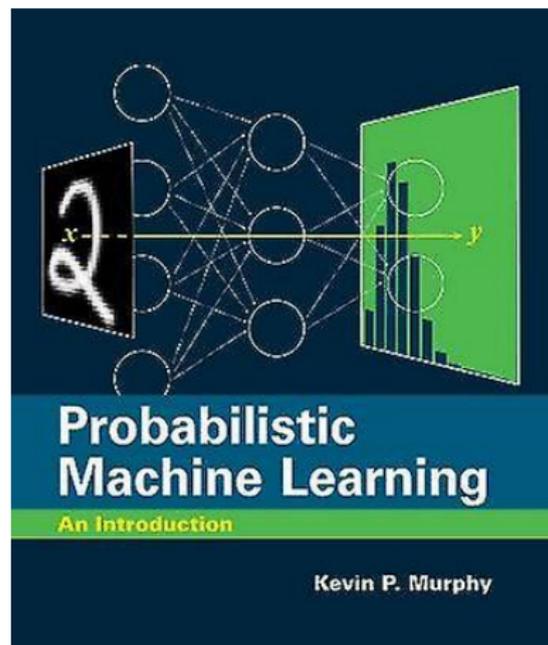
Frequentist vs. Bayesian Statistics

Main differences:

- For a Bayesian, all model parameters are random
- Bayesian inference is conditional on a fixed data set
Frequentists want to repeat the experiment

Questions:

- How does Bayesian inference connect with Statistical/Machine Learning principles?
- How do we effectively teach our DS students the basics of probability theory, (Bayesian) Statistics and Decision Analysis?



K.P. Murphy: Probabilistic Machine Learning. MIT Press 2022, 2023

Most important concepts

Creative Task of Statisticians: Data **Modelling**

Note: "All models are wrong, but some of them are useful"
(George E.P. Box 1978)

Looking into Data Science/ML books: Regression and Classification
(Models) dominate the contents

Basically, the underlying concept for both is the same:

Conditional Expectation $\mathbb{E}[Y|x_1, \dots, x_k] = f(x_1, \dots, x_k)$

Continuous Y : Regression case

Discrete (multinomial) Y : Classification case

In this talk: deal with both central topics

Regression: Linear regression ... Gen. linear (mixed) regression
... Additive regression ... Gaussian Process regression

Classification: Clustering ... Bayes Deep Learning

Bayes's Formula

$$P(H_k|Data) = \frac{P(Data|H_k) \cdot P(H_k)}{\sum_{i=1}^n P(Data|A_i) \cdot P(H_i)}$$

where H_k = Hypotheses (causes, model parameters), $k = 1, \dots, n$

Important in engineering and medicine: **root cause analysis**

Continuous Bayes learning with pdf's rarely known from BSc curricula:

- We start with a model (likelihood) $p(\underline{x}|\theta)$ for the observed data $\underline{x} = (x_1, \dots, x_n)^T$ given a vector of unknown parameters θ
- We add a prior (probability) density $p(\theta)$
- The posterior density of θ is then given by

$$p(\theta|\underline{x}) = \frac{p(\underline{x}|\theta)p(\theta)}{\int_{\Theta} p(\underline{x}|\theta)p(\theta)d\theta} = \frac{p(\underline{x}|\theta)p(\theta)}{p(\underline{x})}$$

where Θ denotes the parameter space

Bayesian Decision Theory

This is often written as

$$\text{posterior} \propto \text{likelihood} * \text{prior}$$

since $p(\underline{x})$ is just a *normalizing constant*.

Bayesian statistical inference (point and interval estimation, hypothesis testing) follow from posterior summaries. For example, the posterior means/medians/modes offer **point (MAP) estimates** of θ , while the quantiles yield **credible intervals**.

Finding an **optimal decision** d requires an evaluation criterion, called **loss function** for decisions (estimates, predictions,...) : $L(\theta, d)$

Bayesian decision principle:

Integrate over Θ to get the **posterior expected loss**

$$E[L(\theta, d)|\underline{x}] = \int_{\Theta} L(\theta, d) * p(\theta|\underline{x})d\theta$$

and minimize w.r.t. d

- Generic packages with R-links:
 - WinBUGS, www.mrc-bsu.cam.ac.uk/bugs/
 - STAN, <https://mc-stan.org>
 - NIMBLE, <https://r-nimble.org>
- More specialized tool:
BayesianTools (Hartig et al. 2017)
can run different MCMC algorithms
- Bayes linear and generalized linear (mixed) models:
 - MCMCpack
 - MCMCglmm
- R package for learning and first steps:
LearnBayes (1- and 2-param. problems)

- New Bayesian tool: **INLA**

van Niekerk, J., Krainski, E., Rustand, D. and Rue, H. (2023). A new avenue for Bayesian inference with INLA. Computational Statistics & Data Analysis, 181, 107692.

- CRAN repository of R, "Task View for Bayesian Inference":
<https://cran.r-project.org/web/views/Bayesian.html>.

Central Role of **Gaussian Process Regression**:

Flexible nonparametric framework based on stochastic processes. Comprehensive introduction is given in the textbook by Rasmussen and Williams (2006).

- as limits of BNN (N. Polson Bayesian Analysis 2017, implicit already in papers/books by R. Neal 1996,2002,...)
- regularization by reference priors for parameters of covariance kernel

Gaussian Process Regression

GP-Regression (Kriging) models replace inflexible RSM models

⇒ response = realization of stochastic process

$$Y(\mathbf{x}) = \underbrace{f(\mathbf{x})} + \underbrace{Z(\mathbf{x})}$$

trend Gaussian Process
(zero mean)

Effect: good approx. over a wide range of different designs and sample sizes and well-defined basis for statistical framework

$$Y(\cdot) \sim GP(f(\mathbf{x}), \sigma^2 R(\cdot, \cdot))$$

Main difference to geostatistical settings:

- \mathbf{x} is no spatial coordinate vector
- usually, higher dimensional settings: $K > 3$

Covariance structure

covariance function: $\text{Cov}(Z(\mathbf{x}_i), Z(\mathbf{x}_j)) = \sigma^2 \underbrace{R(\mathbf{x}_i, \mathbf{x}_j)}_{\text{correlation function}}$
↓
process variance

Common assumptions:

- 1 covariance-stationarity, i.e.

$$R(\mathbf{x}_i, \mathbf{x}_j) = R(\mathbf{x}_i - \mathbf{x}_j)$$

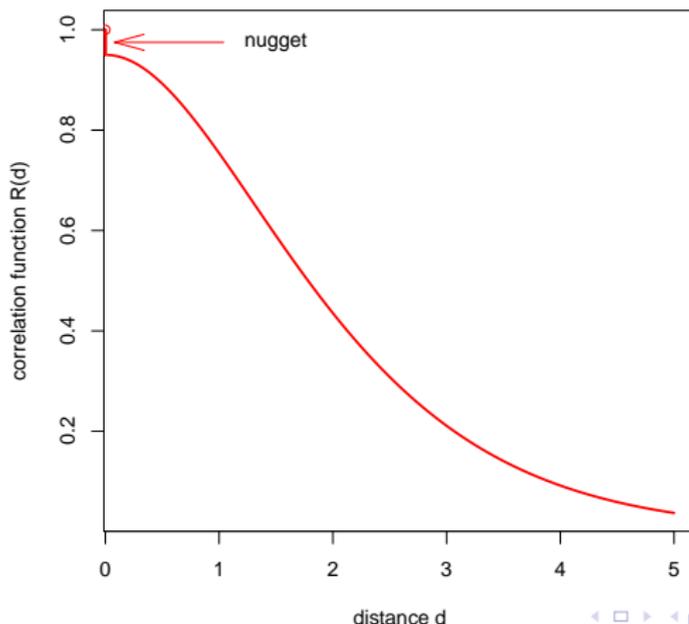
- 2 (tensor-)product correlation structure

$$R(\mathbf{x}_i, \mathbf{x}_j) = \prod_{m=1}^k \underbrace{R_m(|x_{im} - x_{jm}|)}_{\text{univariate Matérn c.f.}}$$

Matérn c.f. $\nu = \frac{5}{2}$

$$R(d) = \left(1 - \frac{\tau^2}{\sigma^2}\right) * \left(1 + \frac{\sqrt{5}d}{\theta} + \frac{5d^2}{3\theta^2}\right) \exp\left(-\frac{\sqrt{5}d}{\theta}\right), \quad d > 0$$

τ^2 = nugget (measurement error)



Aims

- higher flexibility in meta-modelling
- numerical stability: robustness of parameter estimates, esp. for correlation parameters

Solution: Bayesian approach using additive models and (objective) reference priors

Additive model:

$$\mathbb{E}Y(\mathbf{x}) = f_0 + \sum_{i=1}^k f_i(x_i) + \sum_{i < j} f_{ij}(x_i, x_j) + \dots + f_{12\dots k}(x_1, \dots, x_k)$$

Functional ANOVA Representation

Novelty of our recently proposed concept: Combination of AGP with robust reference priors proposed by Gu, Wang and Berger (AS 2018) + new sampling design scheme

Our new model: Second order AGP with

$$f_i \sim N(\mu_i, \sigma^2 R_i)$$

$$f_{ij} \sim N(\mu_{ij}, \sigma^2 R_i R_j)$$

Result: AGP $Y(\mathbf{x}) \sim N(\mu, \sigma^2 R(\cdot, \cdot))$, locally constant trend

$$\text{and } R(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^k R_i(x_i, x'_i) + \sum_{i=1}^k \sum_{j=i+1}^k R_i(x_i, x'_i) R_j(x_j, x'_j) + \delta_{\mathbf{xx}'} \tau^2$$

Profile likelihood approach often fails!

Remedy: robust Bayes prediction using reference priors of the form

$$\pi^R(\mu, \sigma^2, \theta^*) = \frac{\pi^R(\theta^*)}{\sigma^2}$$



correl. parameters

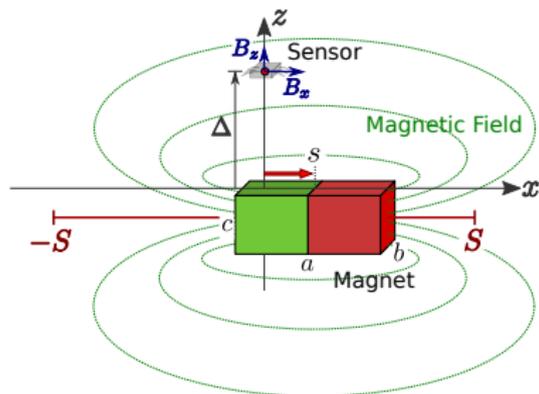
R-implementation fully described in

Vollert, Ortner & Pilz (2019): Robust Additive Gaussian Process Models Using Reference Priors and Cut-Off-Designs, J. Applied Mathematical Modelling 65 (2019), 586-596

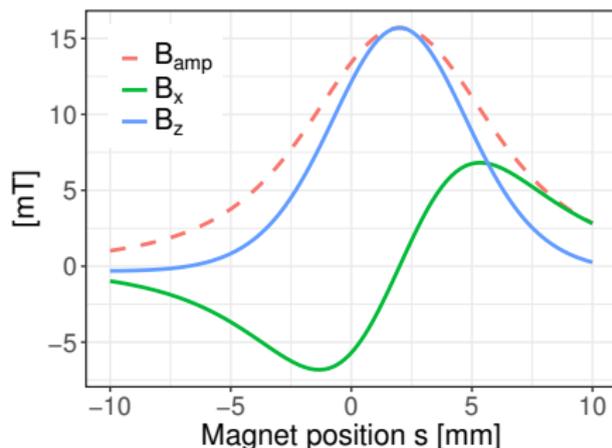
Automotive Application

AGP modelling based on FEMs for
geometric and material parameter optimization problems, e.g.
Magnetic field shaping for position and orientation detection systems

a)



b)



Spike and slab

Intelligent use of Bayesian ideas can provide performance gains, even for the common and still most widely used **multiple linear regression** models.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}; \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

Spike and slab: use mixture priors, with a spike concentrated around zero and a comparably flat slab, to perform variable selection.

In our recent paper

Posch, Arbeiter and Pilz (2020): A novel Bayesian approach for variable selection in linear regression models. CSDA 144 (2020) we use these ideas in combination with variable selection.

Setting is based on a random set $\mathcal{A} \subseteq \{1, \dots, p\}$ that holds the indices of the active predictors

We assign a prior to \mathcal{A} which depends on the cardinality of the set $|\mathcal{A}|$ as well as on the actual elements of \mathcal{A}

Penalized Zellner g -prior

Also, to overcome problems with singularity of $\mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}}$ for $k > n$, we consider a ridge penalized version of the g -prior:

$$\beta_{\mathcal{A}} | g, \sigma^2, \mathbf{X}_{\mathcal{A}} \sim \mathcal{N}(\mathbf{0}, (g^{-1} \sigma^{-2} \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}} + \lambda I_k)^{-1}),$$

with small $\lambda > 0$ and complete hierarchical representation

$$p(\mathcal{A} = \{\alpha_1, \dots, \alpha_k\}) \propto (p_{\alpha_1} + \dots + p_{\alpha_k}) \frac{1}{k} \tilde{p}(k),$$

$$g \sim \text{IG} \left(\frac{1}{2}, \frac{n}{2} \right),$$

$$\sigma^2 \sim \text{IG}(a, b)$$

Our main result: model specifications are **consistent** in terms of model selection:

$$\text{p} \lim_{n \rightarrow \infty} p(M_{\mathcal{A}} | \mathbf{y}, \mathbf{X}) = 1 \quad \text{and} \quad \text{p} \lim_{n \rightarrow \infty} p(M_{\mathcal{A}'} | \mathbf{y}, \mathbf{X}) = 0 \quad \text{for all } \mathcal{A}' \neq \mathcal{A},$$

i.e. the true model will be selected provided we have enough data.

Real-world studies, incl. The diabetes data set (Efron 2004), see R-package *care*

Predictors: age, sex, body mass index, average blood pressure, and six blood serum measurements, measured from $n = 442$ diabetes patients.

Target variable: quantitative measure of disease progression one year after baseline.

Burn-in: 10,000 samples are deleted,

Thinning: every 10-th one is deleted.

For each of the observed Bayesian models 50,000 (dependent) samples are generated

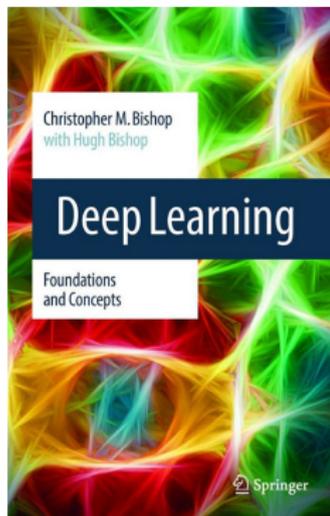
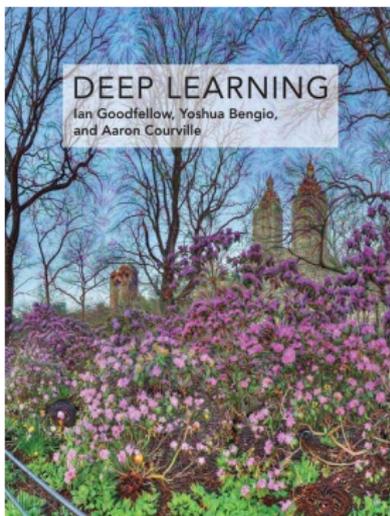
Performing a 5-fold cross-validation, the proposed approach achieves the lowest MMSE as well as the lowest MMAD and thus performs better than all methods under comparison.

Method	MMSE	MMAD
Our approach	0.4873534	0.5678801
Lasso	0.492067	0.571596
Adaptive lasso	0.4939229	0.5736721
Elastic net	0.4922686	0.5706994
Bayesian lasso	0.4924316	0.5736084
Bayesian adaptive lasso	0.4997307	0.5786672
Bayesian elastic net	0.4895844	0.5727555
Horseshoe	0.4903684	0.5711527
Horseshoe+	0.4919946	0.5727804
Spike and slab (VI)	0.5179594	0.5894747
Spike and slab (EM)	0.4893634	0.568348

IV. Bayesian Deep Learning

Popularity of **Deep Learning** is increasing rapidly: excellent results in many fields of applied machine learning, including computer vision and natural language processing

Excellent overview in the monographs of Goodfellow et al. (MIT Press 2016) and Ch.M. Bishop and H. Bishop (Springer 2024)



Note: Deep NNs act as Gaussian Processes, see Lee et al. 2018

Bayesian DL overcomes drawbacks of classical DL:

- Network parameters are treated as random variables
- Uncertainty regarding parameters is directly translated into uncertainty about predictions
- Robustness to overfitting (built-in **regularization**)

We need, however, **ABC** methods to compute **posteriors**

- Laplace approximation (INLA)
- Variational inference, usually with independent Gaussians

Note: Dropout regularization (Gal and Ghahramani 2015) acts like VI, leads to **thinning of NNs**.

Measuring Uncertainty in Deep Neural Networks

Novel approach for training DNNs using Bayesian techniques presented in

K. Posch and J. Pilz: Correlated Parameters to Accurately Measure Uncertainty in Deep Neural Networks. IEEE Transactions on Neural Networks and Learning Systems, Vol. 32 (2021) No. 3, 1037 - 1051

Our novelty comprises

- variational distribution as product of multiple multivariate normals with **tridiagonal** covariance matrices
- correlations are assumed to be identical \Rightarrow only a few additional parameters need to be optimized

Rationale: Dependent tridiagonal (instead of diagonal only)

Gaussians effect an exchange of information between NN layers and neurons

Also, our approach allows an easy evaluation of model uncertainty and is robust to overfitting.

Let \mathbf{W} denote the rv covering all parameters (weights and biases) of a given neural net \mathbf{f} . Further, let $p(\mathbf{w})$ denote the prior regarding \mathbf{W} . According to the Bayes's theorem the **posterior** distribution of \mathbf{W} is given by the density

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{w}, \mathbf{X})p(\mathbf{w})}{\int p(\mathbf{y}|\mathbf{w}, \mathbf{X})p(\mathbf{w}) d\mathbf{w}}$$

where $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_{tr}\}$ denotes a set of training examples and $\mathbf{y} = (y_1, \dots, y_{tr})^T$ holds the corresponding class labels.

The integral above is commonly intractable. **Variational inference** aims at approximating the posterior with the so-called **variational density** $q_\phi(\mathbf{w})$

The variational parameters ϕ are optimized by minimizing

$$KL(q_{\phi}(\mathbf{w})||p(\mathbf{w}|\mathbf{y}, \mathbf{X})) = \mathbb{E}_{q_{\phi}(\mathbf{w})} \left(\ln \frac{q_{\phi}(\mathbf{w})}{p(\mathbf{w}|\mathbf{y}, \mathbf{X})} \right)$$

VI has been used in (B)DL since the invention of Variational AutoEncoders in late 2013 by Kingma and Welling (from the University of Amsterdam).

Since the posterior is unknown this divergence cannot be minimized directly. However, minimization of KL is equivalent to the minimization of the so-called negative log **evidence lower bound**

$$ELBO = -\mathbb{E}_{q_{\phi}(\mathbf{w})} [\ln p(\mathbf{y}|\mathbf{w}, \mathbf{X})] + KL(q_{\phi}(\mathbf{w})||\text{prior } p(\mathbf{w}))$$

Commonly, mini-batch gradient descent is used for optimization.

Note: Bayesian deep learning penalizes deviations of the variational distribution from the prior.

Crucial difference to frequentist DL: **sampled** network parameters.

We have implemented the proposed approach by modifying and extending the popular open-source Deep Learning framework **Caffe** (Jia et al. 2014).

The **Pseudocode** of our implementation is presented in our paper.

Performance evaluation

Comparison includes the frequentist approach, the proposed approach without correlations (see Steinbrener, Posch and Pilz 2020), and, finally, the popular approach which applies dropout before every weight layer in terms of a Bernoulli variational distribution (Gal and Ghahramani 2015).

Criteria: prediction accuracy and quality of the uncertainty information

Modification: Bayes Deep Learning for 3D point cloud segmentation
Application in Automotive Industry, BMW Group Munich-Germany

PyTorch Implementation Two recent publications in MDPI journals

"Entropy 2021" and "Modelling 2021"

Joint work with my youngest PhD Christina Petschnigg:

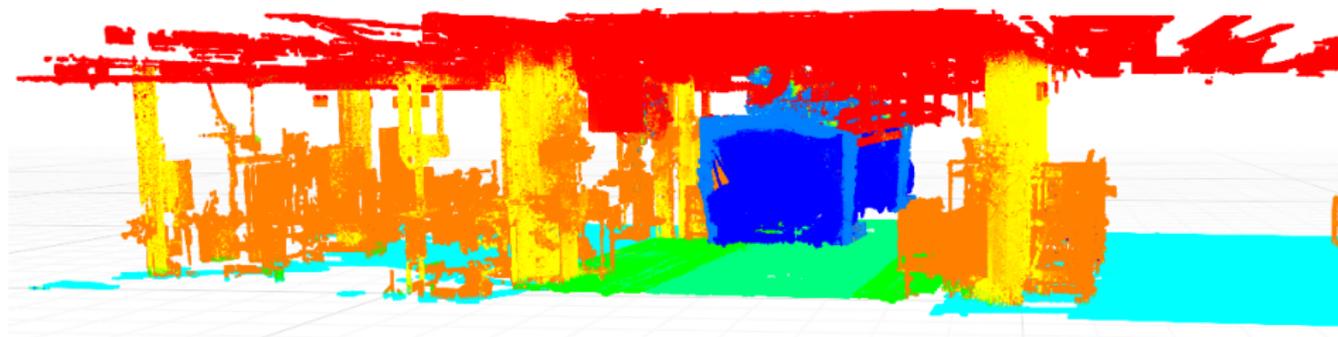
Ch. Petschnigg and J. Pilz: Uncertainty Estimation in Deep Neural Networks for Point Cloud Segmentation in Factory Planning. Modelling 2021, 1, 1-17

Ch. Petschnigg, M. Spitzner, L. Weitzendorf and J. Pilz, From a Point Cloud to a Simulation Model—Bayesian Segmentation and Entropy Based Uncertainty Estimation for 3D Modelling. Entropy 23 (2021) 301, 1 - 27

BMW Assembly Line



Data recording



Classes: Car, Hanger, Floor, Band, Lineside, Wall, Column, Ceiling, Clutter

Please, also have a look at our most recent contributions to **merging** ideas from **Machine Learning** and **Bayesian Statistics**:

A. Jenul, St. Schrunner, J. Pilz and O. Tomic:

A user guided Bayesian framework for ensemble feature selection in life science applications (UBayFS).

Machine Learning (2022) 111:3897 – 3923

K. Posch, Ch. Truden, P. Hungerländer and J. Pilz: A Bayesian approach for predicting food and beverage sales in staff canteens and restaurants.

Int. Journal of Forecasting 38 (2022), 321-338

K. Posch, Ch. Truden, M. Arbeiter, M. Pleschberger and J. Pilz: Variable Selection Using Nearest Neighbor Gaussian Processes. Under Review, 2025 (Preliminary version on arXiv)

References

O. Roustant, D. Ginsbourger, Y. Deville, DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization, *Journal of Statistical Software* 51 (2012) 1, 1–55

H. Kazianka, J. Pilz, Objective Bayesian analysis of spatial data with uncertain nugget and range parameters, *The Canadian Journal of Statistics* 40 (2012), 304–327

M. Gu, X. Wang, J. O. Berger, Robust Gaussian stochastic process emulation, *Annals of Statistics* 46 (2018) 6A, 3038 - 3066

M. Gu, J. Palomo, J.O. Berger, RobustGaSP: Robust Gaussian Stochastic Process Emulation in R. *The R Journal* Vol. 11/01, June 2019

N. Vollert, M. Ortner and J. Pilz, Robust Additive Gaussian Process Models Using Reference Priors and Cut-Off-Designs. *J. Applied Mathematical Modelling* 65 (2019), 586-596

References

- J. Steinbrener, K. Posch and J. Pilz. Variational Inference to Measure Model Uncertainty in Deep Neural Networks. *Sensors* 2020, 20, 6011; doi:10.3390/s20216011
- K. Posch, M. Arbeiter and J. Pilz, A novel Bayesian approach for variable selection in linear regression models. *Computational Statistics and Data Analysis* 144 (2020), 106881
- Alhamzawi, Rahim and Taha Mohammad Ali, Haithem. The Bayesian adaptive lasso regression. *Math. Biosciences* 303(2018), 75 - 82
- R. Tibshirani, Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1996) 1, 267–288
- T. Park and George Casella, The Bayesian Lasso. *Journal of the American Statistical Association* 103 (2008) 482, 681 - 686

References

- Hui Zou, The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* 101 (2006) 476, 1418-1429
- Ch. Leng, Tran, Minh-Ngoc and D. Nott, Bayesian adaptive Lasso. *Annals of the Institute of Statistical Mathematics* 66 (2014) 2, 221 - 244
- Zou, Hui and T. Hastie, Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2005) 2, 301 - 320
- A. Huang, Shizhong Xu and Xiao-hui Cai, Empirical Bayesian elastic net for multiple quantitative trait locus mapping. *Heredity* 114 (2015), 107 - 115
- M. Wang, Sun, Xiaoqian and Lu, Tao, Bayesian structured variable selection in linear regression models. *Computational Statistics* 30 (2015) 1, 205 - 229

References

A. Zellner, On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions. P. K. Goel and A. Zellner, Eds., Basic Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, 1986, 233 - 243

Meili C. Baragatti and D. Pommeret, A study of variable selection using g-prior distribution with ridge parameter. Computational Statistics & Data Analysis 56 (2012), 1920-1934

R.B. Gramacy, monomvn: Estimation for Multivariate Normal and Student-t Data with Monotone Missingness, 2018, R package version 1.9-8, <https://CRAN.R-project.org/package=monomvn>

Rahim Alhamzawi, Brq: An R package for Bayesian Quantile Regression, Working Paper, 2018, <https://cran.r-project.org/web/packages/Brq/Brq.pdf>

J. Friedman, T. Hastie and R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 33 (2010) 1, 1–22

B. Efron, T. Hastie, Trevor, I. Johnstone and R. Tibshirani, Least angle regression. *Ann. Statist.* 32 (2004) 2, 407–499

V. Zuber and K. Strimmer, care: High-Dimensional Regression and CAR Score Variable Selection. R package version 1.1.10, 2017, <https://CRAN.R-project.org/package=care>

J. Lokhorst, B. Venables and B. Turlach, lasso2: L1 constrained estimation aka 'lasso'. R package version 1.2-19, 2014, <https://CRAN.R-project.org/package=lasso2>

V. Ročková and E. I. George, The Spike-and-Slab LASSO. *Journal of the American Statistical Association* 113 (2018) 521, 431 - 444

Anhui Huang and Dianting Liu, EBglmnet: Empirical Bayesian Lasso and Elastic Net Methods for Generalized Linear Models. R package version 4.1, 2016, <https://CRAN.R-project.org/package=EBglmnet>

- C. M. Carvalho, N. G. Polson and J. G. Scott, The horseshoe estimator for sparse signals. *Biometrika* 97 (2010) 2, 465-480
- E. Makalic and D. Schmidt, High-Dimensional Bayesian Regularised Regression with the BayesReg Package. arXiv:1611.06649v3, 2016
- N. G. Polson and J. G. Scott, Local shrinkage rules, Levy processes and regularized regression. *Journal of the Royal Statistical Society (Series B)* 74 (2012) 2, 287-311
- A. Bhadra, J. Datta, N.G. Polson and B. Willard, Brandon, The Horseshoe+ Estimator of Ultra-Sparse Signals. *Bayesian Analysis* 12 (2017) 4, 1105–1131
- P. Carbonetto and M. Stephens, Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. *Bayesian Analysis* 7 (2012) 1, 73–108

References

V. Ročková and E. I. George, EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association* 109 (2014) 506, 828-846

A. Bhattacharya, D.Pati, N. S. Pillai and D. B. Dunson, Dirichlet–Laplace Priors for Optimal Shrinkage. *Journal of the American Statistical Association* 110 (2015) 512, 1479-1490

Zhang, Yan and H. D. Bondell, Variable Selection via Penalized Credible Regions with Dirichlet–Laplace Global-Local Shrinkage Priors. *Bayesian Analysis* 13 (2018), 823–844

Chen, Su and S. G. Walker, Fast Bayesian variable selection for high dimensional linear models: Marginal solo spike and slab priors. *Electronic Journal of Statistics* 13 (2019) 1, 284–309

N.G. Polson and S.L. Scott, Data augmentation for support vector machines. *Bayesian Analysis* 6 (2011) 1, 1–23

References

Zhou, Quan and Y. Guan, Fast Model-Fitting of Bayesian Variable Selection Regression Using the Iterative Complex Factorization Algorithm. *Bayesian Analysis* 14 (2019) 2, 573–594

Ch. Petschnigg and J. Pilz: Uncertainty Estimation in Deep Neural Networks for Point Cloud Segmentation in Factory Planning. *Modelling* 2021, 1, 1-17

J. R. Hershey and P. A. Olsen, Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models. *IEEE International Conference on Acoustics, Speech and Signal Processing* 4 (2007), 317–320

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell, Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv:1408.5093*, 2014

A. Krizhevsky, I. Sutskever, Ilya and G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems* 25 (2012), 1097-1105

References

- Y. Gal, Yarin and Z. Ghahramani, Zoubin, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. Proceedings of The 33rd International Conference on Machine Learning 2015, 1050–1059
- L. Deng, The MNIST Database of Handwritten Digit Images for Machine Learning Research. IEEE Signal Processing Magazine 29 (2012), 141-142
- A. Krizhevsky, Learning Multiple Layers of Features from Tiny Images. 2009
- H. M. D. Kabir, A. Khosravi, M. A. Hosen and S. Nahavandi, Neural Network-Based Uncertainty Quantification: A Survey of Methodologies and Applications. IEEE Access 6 (2018), 36218-36234
- Ch. Petschnigg, M. Spitzner, L. Weitzendorf and J. Pilz, From a Point Cloud to a Simulation Model—Bayesian Segmentation and Entropy Based Uncertainty Estimation for 3D Modelling. Entropy 23 (2021) 301, 1 - 27

van Niekerk, J. and Rue, H. (2021). Correcting the Laplace method with variational Bayes. arXiv preprint 2111.12945.

Wang, X., Yue, Y. and Faraway, J. J. (2018). Bayesian regression modeling with INLA. New York: Chapman & Hall/CRC.