The importance of bayesian statistics and decision theory in modern data science education

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 26-28, 2023 / Frankfurt, Germany

I. Introduction

Over the last decade, many conferences under the heading of "Machine Learning/Deep Learning", "Data Science", "Data Analysis/Analytics", "Big Data", but much fewer with joint theme "**Statistics** and Data Science", "**Statistics** and Data Analytics", and still fewer with sole theme "(Applied) Statistics".

Good message: Joint theme conferences (research monographs, textbooks, ...) and university curricula of Data Science study programs incl. sound education in statistics are gaining ground!

Main Drivers of recent developments/ trends have been

- Availability of massive data sets
- Modern Computer Technology and Computing Environments

Essential role of Probability Theory, Information Theory and Statistics is getting more and more acknowledged! But, we need to double our efforts to propagate the underlying probabilistic and statistical basis of Data Science and and our contributions to it!

Introduction

"Statistics is the grammar of science." Karl Pearson (1892)

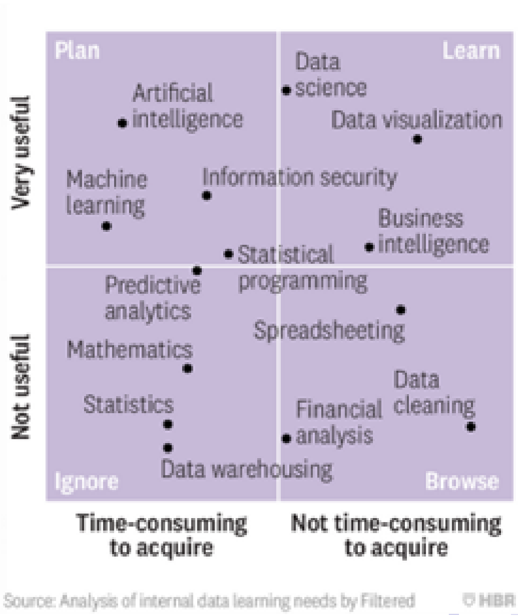
"Those who ignore statistics are condemned to reinvent it. Statistics is the science of learning from experience." Bradley Efron (2006)

"Data can tell lies. – Big Data can tell bigger lies. – The big thing for small data is random error. – The big thing for big data is bias."
Chris Wild (2017)

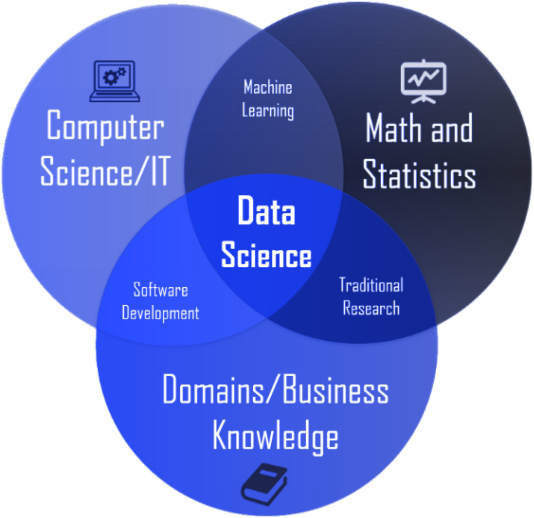
Fundamental ideas in statistics: uncertainty and variation.
Two of these key developments over the last decades are **bootstrapping** (Bradley Efron, 1979) and **Monte Carlo Markov Chain** (MCMC, Gelfand and Smith 1990) methods, which make it possible to compute large hierarchical models, e.g. in Bayesian statistics, computational physics and chemistry, computational biology and linguistics, etc.

The widespread use of such powerful computational tools would have been impossible without the emergence of the statistical programming language R (released in 1993)

A Very Dangerous Data Science Article

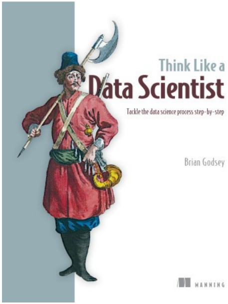


What is Data Science?



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



However, it has been a long race! When I started teaching Bayesian Statistics and Decision Analysis in 1979 after having obtained a PhD with a thesis on Bayesian regression estimation and design the academic scene in statistics was heavily dominated by frequentist statistics with the exception of some UK universities (London, Nottingham, Warwick) and US universities (UCLA, Purdue, Ohio SU, Minnesota, Duke).
In Central Europe: some Italian and French Bayesian statisticians (de Finetti, Christian Robert)

Bayesian Origins

Good historic account in the paper by
S.E. Fienberg: When Did Bayesian Inference Become "Bayesian"?
Bayesian Analysis Vol 1 (2006)

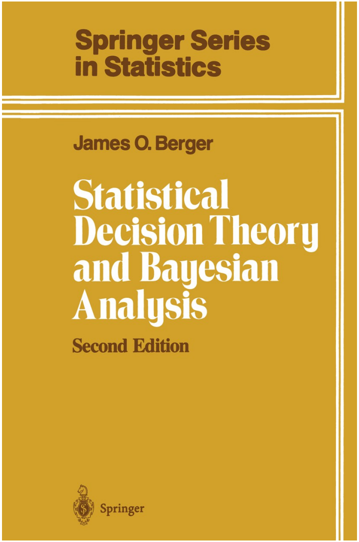
Raiffa, H. and Schlaifer, R. (1961). Applied Statistical Decision Theory. Division of Research Graduate School of Business Administration, Harvard University.

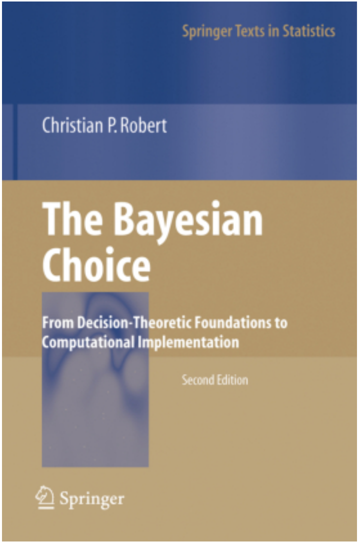
DeGroot, M. A. (1970). Optimal Statistical Decisions. McGraw-Hill.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. New York: Wiley.

J.O. Berger: Statistical Decision Theory. Springer 1985

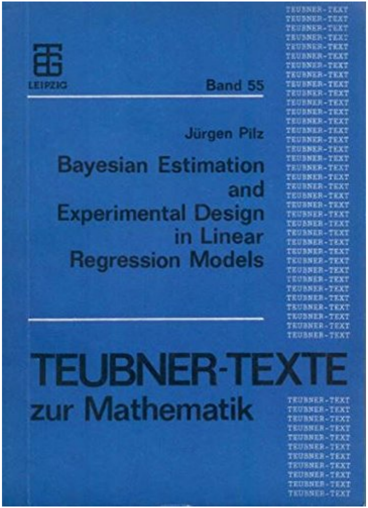
Ch. Robert: The Bayesian Choice. Springer 1997

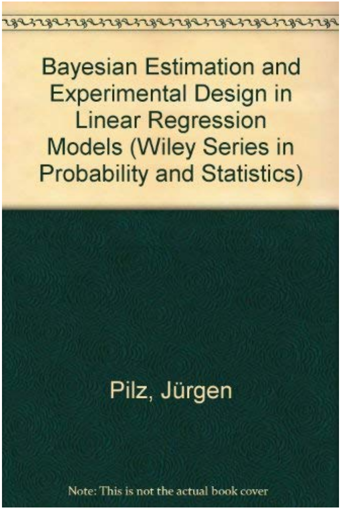




J.O. Berger (left): Springer 1985, Ch.P. Robert (right): Springer 1997

My own modest early contributions





J. Pilz: ... Teubner ed. (left) 1983, ext. lic. ed. by Wiley (right) 1991

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?

Probabilistic Machine Learning

Ideal starting point for planning effective DS Curriculum:

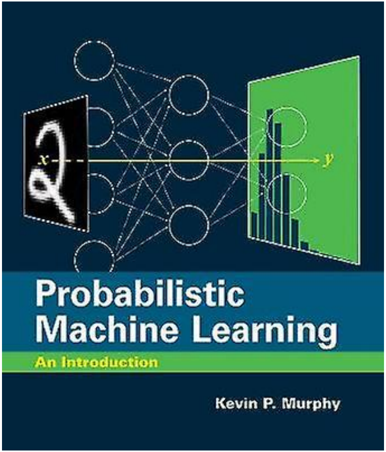
Ghahramani, Z. (2015): Probabilistic machine learning and artificial intelligence. Nature 521:452-459

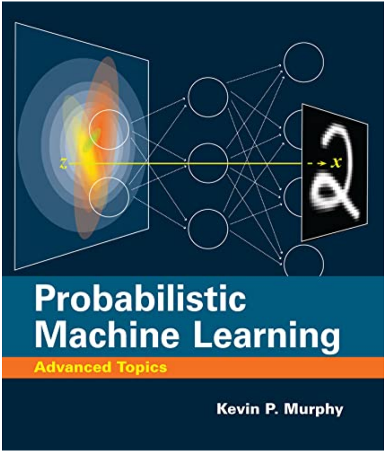
- Probabilistic modelling provides a framework for understanding what **learning** is
- Probabilistic framework describes how to represent and manipulate **uncertainty** about models and predictions
- PM plays a central role in scientific data analysis, machine learning, robotics, cognitive science, and artificial intelligence.

Article provides an introduction to this probabilistic framework, and reviews some state-of-the-art advances in the field, namely:

- Probabilistic programming, Bayesian optimisation
- Data compression, and automatic model discovery.

Perfect textbook





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 $\mathbb{E}[Y|x_1, \ldots, x_k] \; = \; f(x_1, \ldots, x_k)$

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

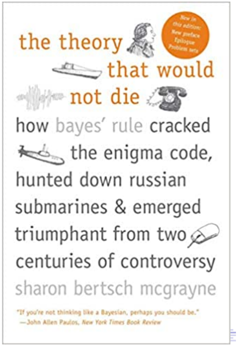
Bayes's Formula

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

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

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

For **further applications** in art sciences, literature, music, ... see



Bayes's Formula

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 $\underline{x} = (x_1, \ldots, x_n)^T$ 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
- This is often written as

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

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

Bayesian Decision Theory

Probabilistic treatment is not only a technical aspect,
it provides a **lot of advantages**

- The Bayesian approach expands the class of models and easily handles settings that are precluded in classical settings: **Regularization** of ill-posed problem settings
- Moreover, we have **complete class theorems** from **Statistical Decision Theory** stating that, for any non-Bayesian decision (estimator, test, predictor), there is a Bayes decision which is at least as good (usually even better).
- (Maximum) likelihood results are often obtained as limiting Bayesian results w.r.t. vague or **non-informative priors**

Bayesian Decision Theory

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

Software for Bayesian Inference

- Generic packages with R-links:
  - WinBUGS, www.mrc-bsu.cam.ac.uk/bugs/
  - STAN, https://mc-stan.org
  - NIMBLE, https://r-nimble.org
- More recently developed 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)
- CRAN repository of R, "Task View for Bayesian Inference":
  https://cran.r-project.org/web/views/Bayesian.html.

Software for Bayesian Inference

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

New Bayesian tool: **INLA** (well-known from Bayesian spatial statistics)

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

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.

Teaching Problems

Within (university) Master curricula in Mathematics you have plenty of time to make a deep dive into the many facets of Bayesian Statistics:

- Bayesian Multivariate Statistics, Bayesian Time Series Analysis Bayesian Decision Analysis
- Bayes linear, generalized linear and additive models
- Bayesian Computing
- ....

Reason: You can build on sufficient basics knowledge in Maths/Probability Theory and Statistics from BSc curriculum

Remark: During my university career, spanning more than 40 years now, I have designed and taught $>$ 30 different statistics courses, incl. an early course on "Neural Networks" in 1998, based on Brian Ripley's R-package **nnet** and Timothy Master's book "Advanced Algorithms for Neural Networks - A C++ Sourcebook", Wiley 1995

Starting from BSc in Engineering and Economics

How to manage teaching the essential elements of modern Bayesian Probability and Statistics in a Master course "Applied Data Science", e.g. such as that at
Carinthian University of Applied Science in Villach, Austria, which started in Fall 2021

https://www.fh-kaernten.at/en/study-program/engineering-it/master/applied-data-science

Information and Probability Theory (Module 1.1)

- Combinatorics and enumeration: permutations, variations and combinations with and without replication
- Different definitions of probability: classical approach (Laplace), axiomatic approach (Kolmogorov), relative frequency
- Conditional and total probability, independence, inverse probability (Bayes Formula), random variables (rv's)
- Basic discrete distributions: Binomial, Poisson, geometric, NegBinomial Distribution

Bayesian Statistics and Advanced Topics

- Basic continous distributions: uniform, Gaussian, Student-t-, Exponential, Gamma, Beta, Weibull and Lognormal d.
- Characteristic quantities and functions of probability distributions: mean, variance, skewness, kurtosis, p-quantile, moment generating function, Jacobi-transformation, law of large numbers and central limit theorem
- Multivariate distributions: multinomial, multivariate Gaussian, Dirichlet d., correlations and covariances, transformations (convolution and ratio of rv's) and Jacobian matrix
- Introduction into information theory: binary entropy, Hamming code, entropy and relative entropy (Kullback-Leibler divergence), conditional entropy and differential entropy
- (Shannon-) information measures: conditional and mutual (multivariate) information

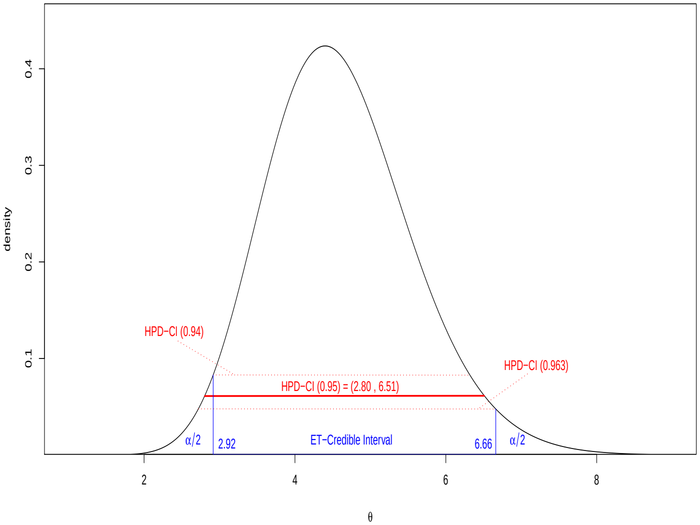
Module Bayesian Statistics

Bayesian Statistics (within Module 1.2 Statistics)

- Bayes rule for probability densities
- Conjugate priors, Bayes estimates
- Predictive distributions
- Bayes credible regions
- Bayesian software (LearnBayes, BayesianTools)

Start with motivating example: (exteremely) rare events,
where classical Maximum Likelihood approaches collapse.

Bayes Credible Intervals



Equal-tailed (blue) and HPD- Credible (red) Intervals

Module Advanced Topics

Advanced Topics: Markov Chain Modeling, Time Series Modeling and Analysis, Intro to Bayesian Deep Learning Models (Module 7.1)

- Markov Chains: transition matrix, recurrent and transient states, ergodic distributions, steady states, first passage time and recurrence time, random walks and other applications
- Monte-Carlo Markov Chain methods for Bayes statistical computing: Gibbs sampling, Metropolis-Hastings- Algorithm
- Time series modeling: white noise, autoregression, filtered series, trend and auto-covariance function, moving average models, stationarity, differencing, ARIMA models, seasonality, Holt-Winters smoothing, R- and Python libraries for statistical and deep learning time series analysis
- Bayes linear and generalized linear regression modeling
- Gaussian Processes (GP) for Regression and Classification: treed Gaussian processes, stationary and non-stationary Gaussian processes, covariance functions, surrogate models, additive GP, deep Gaussian process modeling

Module Advanced Topics

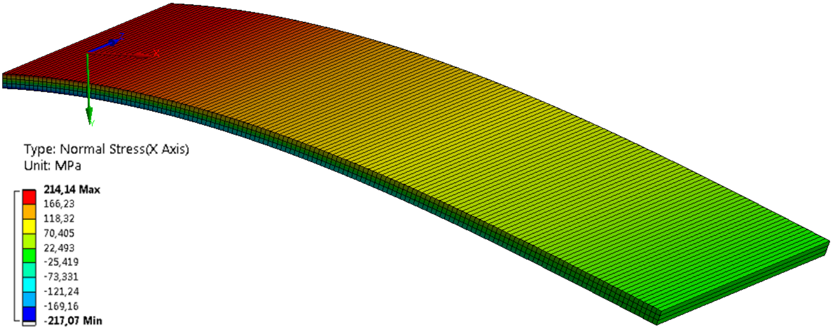
- Decision Analysis, Loss functions, Risk notions, Bayes risk
- Regularization methods for Deep Learning: Bagging and other Ensemble methods, Early stopping, Dropout, Data augmentation, Sparse Deep NNs, Bayesian NNs
- Bayes Deep Learning Models: Bayesian Classification, Kernel methods, NN architectures for Image and 3D-Point Cloud segmentation (LeNet, GoogLeNet, ..., PointNet), Uncertainty Evaluation in (Bayesian) DNNs

Central Role of Gaussian Process Regression:
Flexible nonparametric framework based on stochastic processes.
Comprehensive introduction is given in the textbook by Rasmussen, and Williams.

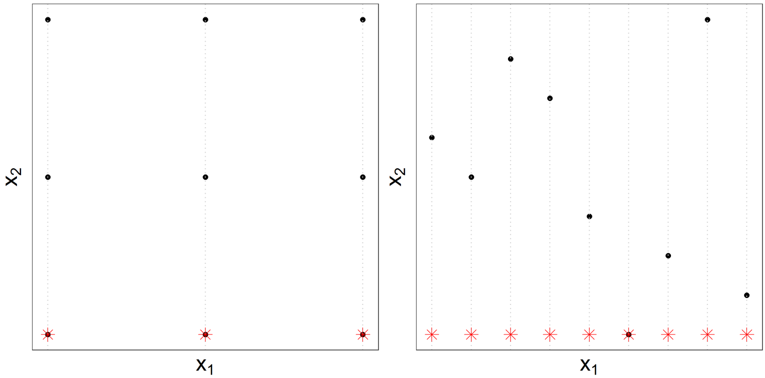
- as limits of BNN (N. Polson Bayesian Analysis 2017, implicit already in papers/books by R. Neal 1996,2002,...)
- regularization by reference priors for parameters of covariance kernel
- space-filling designs: suitable modifications of **LHD**)

II. Gaussian Process Regression

Start with specific application:
Stress testing in semiconductor processing for **thin wafers** (thickness$\leq 40\mu m$)
Kriging metamodel for stress prediction validated against Ramann spectroscopy measurements, FEM simulations
+ Modelling of electrical parameters (signals)



Gaussian Process Regression



Regular (left) and latin hypercube design (right)

Gaussian Process Regression

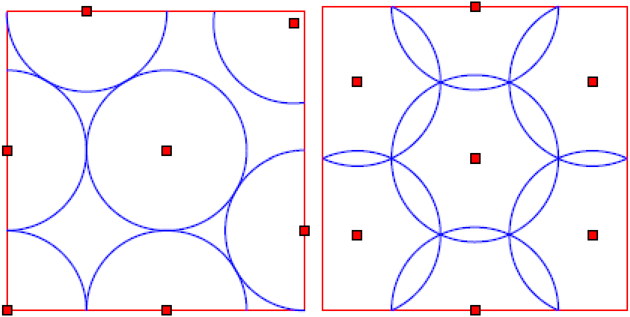
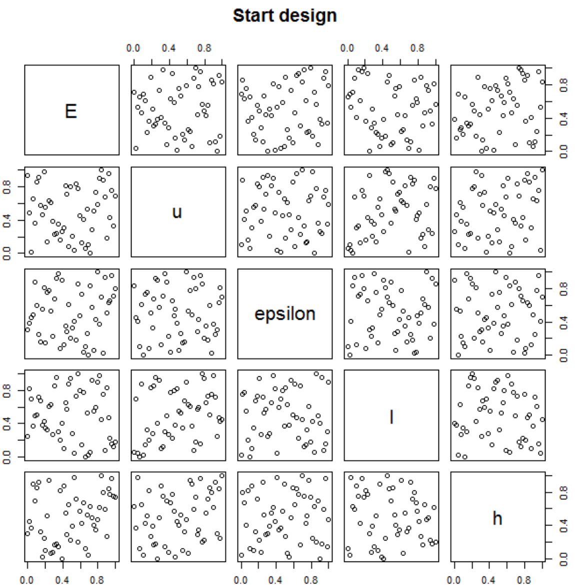
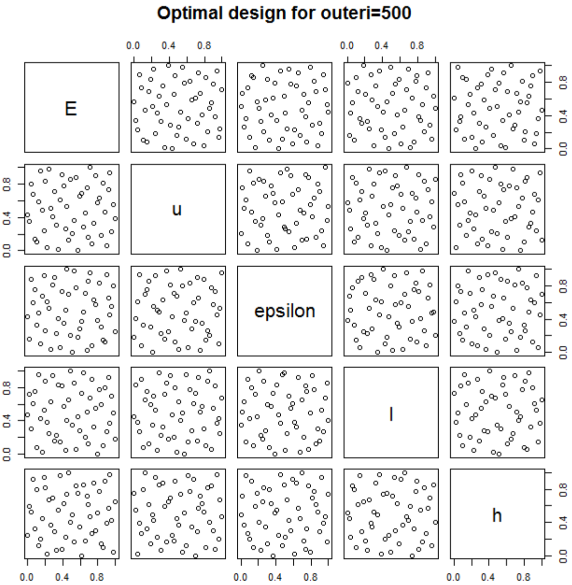


Fig.. Maximin (left) and Minimax (right) designs

Gaussian Process Regression

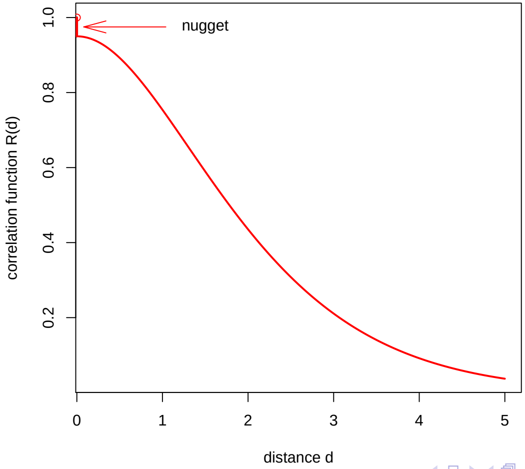


Gaussian Process Regression



Matérn c.f. $\nu = \frac{5}{2}$

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



Robust Additive Gaussian Processes

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

Side effect: high-dimensional optimization problems reduced to a few sub-routines of ≤ 3 dimensions

Additive model:

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

Functional ANOVA Representation

Robust Additive Gaussian Processes

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 Kriging 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

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

Robust Additive Gaussian Processes

Profile likelihood approach often fails!

Remedy: robust Bayes prediction using reference priors of the form

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

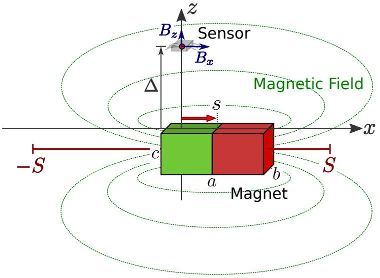
$\theta^* \downarrow$ correl. parameters

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

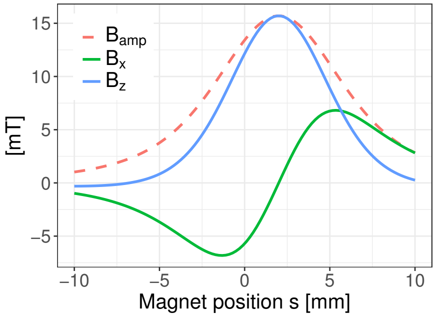
Automotive Application

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.

These Bayesian methods belong to the so-called **spike and slab** approaches: use mixture priors, with a spike concentrated around zero and a comparably flat slab, to perform variable selection.

Note: spike and slab priors are also applied apart from classical regression approaches. E.g. Polson (2011) used this type of priors to regularize support vector machines.

In our recent paper Posch, Arbeiter and Pilz (2020): use of these ideas in combination with variable selection, setting is based on a random set $\mathcal{A} \subseteq \{1, ..., p\}$ that holds the indices of the active predictors, i.e. the predictors with coefficients different from zero.
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:

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

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

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

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

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

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

$$\operatorname*{p\,lim}_{n\to\infty} p(M_\mathcal{A}|\mathbf{y},\mathbf{X}) = 1 \quad \text{and} \quad \operatorname*{p\,lim}_{n\to\infty} p(M_{\mathcal{A}'}|\mathbf{y},\mathbf{X}) = 0 \quad \text{for all } \mathcal{A}' \neq \mathcal{A},$$

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

Data sets

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.

MMSE and MMAD

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 Goodfellow, Bengio and Courville:
Deep Learning. MIT Press 2016

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
- Variational inference, usually with independent Gaussians

Note: Dropout regularization (Gal and Ghahramani 2015) acts like VI

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

Prediction uncertainty

Note: Variational Bayes is just a specific case of local α-divergence minimization:
α-divergence between two densities $p(\mathbf{w})$ and $q(\mathbf{w})$ is given by

$$D_\alpha(p(\mathbf{w})||q(\mathbf{w})) = \frac{1}{\alpha(1-\alpha)}\left(1 - \int p(\mathbf{w})^\alpha q(\mathbf{w})^{(1-\alpha)}\, d\mathbf{w}\right)$$

α-divergence converges for $\alpha \to 0$ to the Kullback-Leibler (KL) divergence typically used in variational Bayes. We used a product of tridiagonal Gaussians as variational density $q(\cdot)$. Moreover, note that

Prediction uncertainty = Epistemic (model) uncertainty
+ Aleatoric (observational) uncertainty

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

KL divergence

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

where $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_\beta\}$ denotes a set of training examples and $\mathbf{y} = (y_1, ..., y_\beta)^T$ holds the corresponding class labels.
Note that $p(\mathbf{y}|\mathbf{w},\mathbf{X}) = \prod_{i=1}^{\beta} \mathbf{f}(\mathbf{x}_i;\mathbf{w})_{y_i}$

The integral above is commonly intractable due to its high dimension β. 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 the KL divergence

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

Since the posterior is unknown this divergence cannot be minimized directly. However, minimization of D_{KL} is equivalent to the minimization of so-called

ELBO

negative log **evidence lower bound**

$$L_{VI} = -\mathbb{E}_{q_\phi(\mathbf{w})}\left[\ln p(\mathbf{y}|\mathbf{w},\mathbf{X})\right] + D_{KL}(q_\phi(\mathbf{w})||\mathit{prior}\ p(\mathbf{w}))$$

Commonly, mini-batch gradient descent is used for optimization

Summing up:

- Frequentist deep learning penalizes (Euclidean) norm of **w**
- Bayesian deep learning penalizes deviations of the variational distribution from the prior.
  Crucial difference: **sampled** network parameters.

Tridiagonal Gaussians

Choosing

$$\mathbf{L}_j := \begin{pmatrix} a_{j1} & & & & \\ c_{j1} & a_{j2} & & & \\ & c_{j2} & a_{j3} & & \\ & & \ddots & \ddots & \\ & & & c_{j,K_j-1} & a_{jK_j} \end{pmatrix}$$

we end up with a **tridiagonal** cov. matrix Σ_j with **equal correlations**

To train a neural net $\mathbf{f}(\cdot, \mathbf{w})$ we need the partial derivatives of the approximation $\widehat{L}_{VI}$ with respect to all variational parameters.

Note: Loss function L = negative log likelihood of the data, i.e.

- L = cross-entropy loss in case of classification and
- L = Euclidean loss in case of regression.

Implementation

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

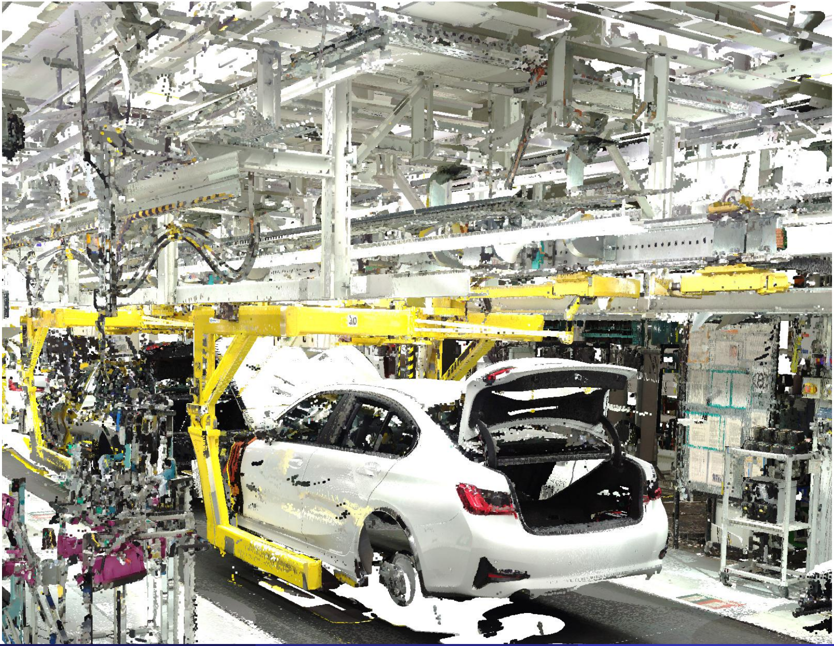
3D Point Clouds

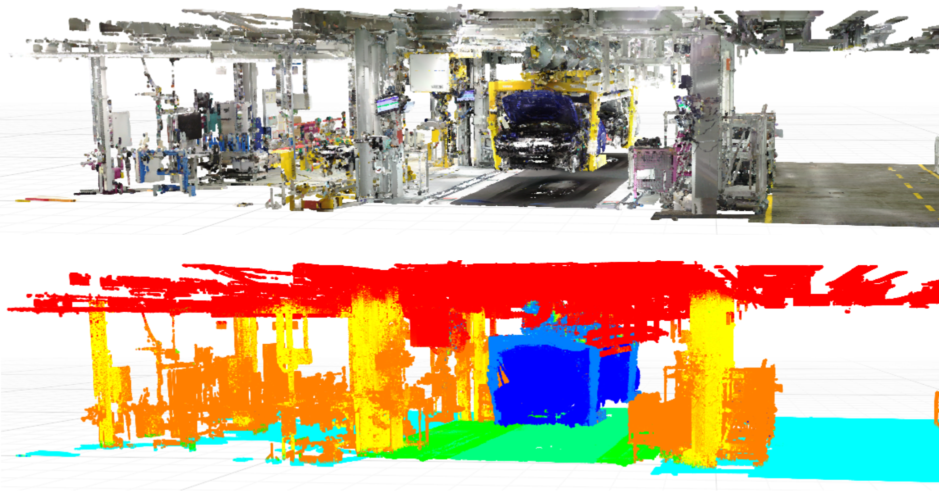
Modification: Bayes Deep Learning for 3D point cloud segmentation
Application in Automotive Industry, BMW Group Munich-Germany

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





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

Concluding Notes

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

Anna Jenul, Stefan Schrunner, Jürgen Pilz and Oliver Tomic:
A user-guided Bayesian framework for ensemble feature selection in life science applications (UBayFS).
Machine Learning (2022) 111:3897 – 3923

Konstantin Posch, Christian Truden, Philipp Hungerländer and Jürgen Pilz: A Bayesian approach for predicting food and beverage sales in staff canteens and restaurants.
Int. Journal of Forecasting 38 (2022), 321-338

Importance of information-theoretic concepts for approximating the posterior distribution!

New promising research avenue:
Generalized Variational Inference

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

Meïli 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

References

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

References

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