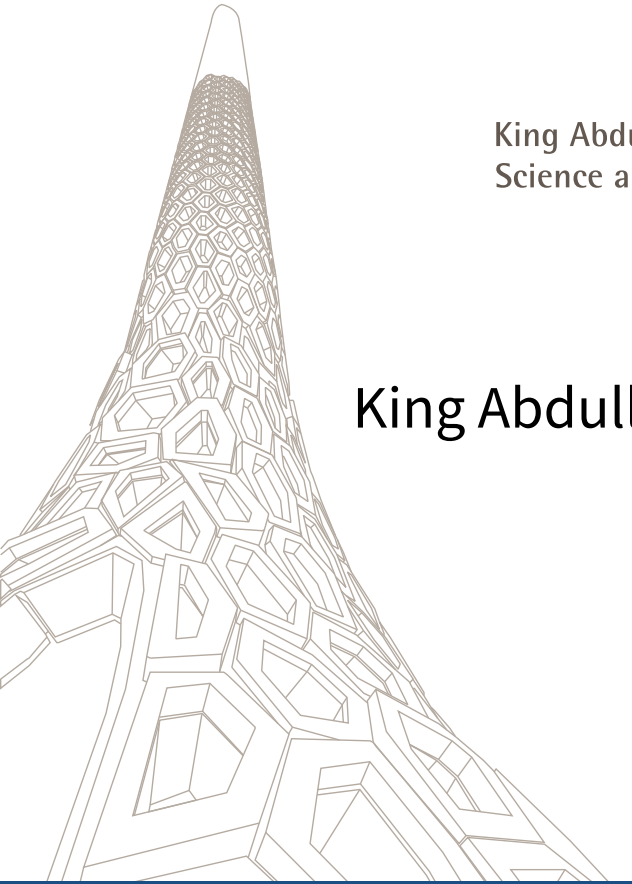
Discussion of “Improper models for data analysis”



King Abdullah University of Science and Technology

جامعة الملك عبدالله للعلوم والتقنية

Håvard Rue
King Abdullah University of Science and Technology
Saudi Arabia

Sep 2022

General comments

- I agree with the questions raised
- "Robustness" is more important that one get impression of today
- We are also towards this, by robustifying Gaussian processes with the aim of this ending up in R-INLA

General comments

- I agree with the questions raised
- “Robustness” is more important that one get impression of today
- We are also towards this, by robustifying Gaussian processes with the aim of this ending up in R-INLA

General comments

- I agree with the questions raised
- “Robustness” is more important that one get impression of today
- We are also towards this, by robustifying Gaussian processes with the aim of this ending up in R-INLA

Improper

- 'Improper' is not a feature, to me it's a shortcut for trying to get away with something annoying.
- If the prior for the intercept is constant, one usually get away with it, but if not (like the posterior is "improper"), the user have an issue.

Improper

- 'Improper' is not a feature, to me it's a shortcut for trying to get away with something annoying.
- If the prior for the intercept is constant, one usually get away with it, but if not (like the posterior is "improper"), the user have an issue.

Improper

- Tukey loss and Gaussian only differ in the tail, so its tempting to interpret this loss as a robustification of the Gaussian instead of a generic loss.
- With this interpretation, then we can modify $L(z) = \min\{1, z^2\}$ into

$$L(z) = \min\{1 + 2\delta(|z| - 1)_+, z^2\}$$

  for some fixed $\delta > 0$.
- Now we are back to model selection and standard theory.
- Tukey's loss will give practical issues, like searching in the dark.
- I do not enjoy non-convex and/or highly multi modal optimisation problems, where (in general) "nothing" will work reliable in practice.

Improper

- Tukey loss and Gaussian only differ in the tail, so its tempting to interpret this loss as a robustification of the Gaussian instead of a generic loss.
- With this interpretation, then we can modify $L(z) = \min\{1, z^2\}$ into

$$L(z) = \min\{1 + 2\delta(|z| - 1)_+, z^2\}$$

  for some fixed $\delta > 0$.
- Now we are back to model selection and standard theory.
- Tukey's loss will give practical issues, like searching in the dark.
- I do not enjoy non-convex and/or highly multi modal optimisation problems, where (in general) "nothing" will work reliable in practice.

Improper

- Tukey loss and Gaussian only differ in the tail, so its tempting to interpret this loss as a robustification of the Gaussian instead of a generic loss.
- With this interpretation, then we can modify $L(z) = \min\{1, z^2\}$ into

$$L(z) = \min\{1 + 2\delta(|z| - 1)_+, z^2\}$$

  for some fixed $\delta > 0$.
- Now we are back to model selection and standard theory.
- Tukey's loss will give practical issues, like searching in the dark.
- I do not enjoy non-convex and/or highly multi modal optimisation problems, where (in general) "nothing" will work reliable in practice.

Improper

- Tukey loss and Gaussian only differ in the tail, so its tempting to interpret this loss as a robustification of the Gaussian instead of a generic loss.
- With this interpretation, then we can modify $L(z) = \min\{1, z^2\}$ into

$$L(z) = \min\{1 + 2\delta(|z| - 1)_+, z^2\}$$

  for some fixed $\delta > 0$.
- Now we are back to model selection and standard theory.
- Tukey's loss will give practical issues, like searching in the dark.
- I do not enjoy non-convex and/or highly multi modal optimisation problems, where (in general) "nothing" will work reliable in practice.

Improper

- Tukey loss and Gaussian only differ in the tail, so its tempting to interpret this loss as a robustification of the Gaussian instead of a generic loss.
- With this interpretation, then we can modify $L(z) = \min\{1, z^2\}$ into

$$L(z) = \min\{1 + 2\delta(|z| - 1)_+, z^2\}$$

  for some fixed $\delta > 0$.
- Now we are back to model selection and standard theory.
- Tukey's loss will give practical issues, like searching in the dark.
- I do not enjoy non-convex and/or highly multi modal optimisation problems, where (in general) "nothing" will work reliable in practice.

Robustness

- If the likelihood is ($\approx$) known, use it
- There are ways to robustify likelihoods, which is case specific
- If there is *no likelihood*, just a "loss function", one can/have-to use it.
- If $L()$ is a loss then also $2 \times L()$ is a loss, hence we run into a "think of a number-game, which mainly impact the "variance".
- We cannot just "estimate" this number the usual way with

$$\pi(y|\tau, ...) \propto \exp(-\tau L())\sqrt{\tau}, \quad \tau \sim \pi(\tau)$$

Robustness

- If the likelihood is ($\approx$) known, use it
- There are ways to robustify likelihoods, which is case specific
- If there is *no likelihood*, just a "loss function", one can/have-to use it.
- If $L()$ is a loss then also $2 \times L()$ is a loss, hence we run into a "think of a number-game, which mainly impact the "variance".
- We cannot just "estimate" this number the usual way with

$$\pi(y|\tau, ...) \propto \exp(-\tau L())\sqrt{\tau}, \quad \tau \sim \pi(\tau)$$

Robustness

- If the likelihood is ($\approx$) known, use it
- There are ways to robustify likelihoods, which is case specific
- If there is *no likelihood*, just a "loss function", one can/have-to use it.
- If $L()$ is a loss then also $2 \times L()$ is a loss, hence we run into a "think of a number-game, which mainly impact the "variance".
- We cannot just "estimate" this number the usual way with

$$\pi(y|\tau, ...) \propto \exp(-\tau L())\sqrt{\tau}, \quad \tau \sim \pi(\tau)$$

Robustness

- If the likelihood is ($\approx$) known, use it
- There are ways to robustify likelihoods, which is case specific
- If there is *no likelihood*, just a "loss function", one can/have-to use it.
- If $L()$ is a loss then also $2 \times L()$ is a loss, hence we run into a "think of a number-game, which mainly impact the "variance".
- We cannot just "estimate" this number the usual way with

$$\pi(y|\tau, ...) \propto \exp(-\tau L())\sqrt{\tau}, \quad \tau \sim \pi(\tau)$$

Robustness

- If the likelihood is ($\approx$) known, use it
- There are ways to robustify likelihoods, which is case specific
- If there is *no likelihood*, just a "loss function", one can/have-to use it.
- If $L()$ is a loss then also $2 \times L()$ is a loss, hence we run into a "think of a number-game, which mainly impact the "variance".
- We cannot just "estimate" this number the usual way with

$$\pi(y|\tau, ...) \propto \exp(-\tau L())\sqrt{\tau}, \quad \tau \sim \pi(\tau)$$

Example: Quantile regression

- Quantile regression is usually *defined* as

> Based on a random sample $\mathcal{D} = \{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$ of $(Y, \mathbf{X})$, the unknown parameters $\boldsymbol{\beta}(\tau)$ can be estimated by $\hat{\boldsymbol{\beta}}(\tau)$, which minimises
>
> $$R_n(\boldsymbol{\beta}, \mathcal{D}) = \sum_{i=1}^{n} \rho_\tau \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right), \tag{2.2}$$
>
> where $\rho_\tau(\mu) = \mu\{\tau - I(\mu < 0)\}^*$ is the quantile loss function given in Koenker & Bassett (1978). In the rest of the paper, we omit the τ in various expressions such as $\boldsymbol{\beta}(\tau)$ for the sake of simplicity.

- This is very weird to me, as this is done even if the likelihood is known.
- This is like always using ordinary least squares instead of GLM's.
- A very incoherent way of thinking
- The *better* way is to use the quantile function[1] which can also be justified for some discrete distr. (Padellini & R, 2018) like Poisson/Binomial/NegBinomial.

[1] *Parametric quantile regression based on the generalized gamma distribution*, by A. Noufaily, M.C. Jones, JRSS-C, 2013)

Example: Quantile regression

- Quantile regression is usually *defined* as

> Based on a random sample $\mathcal{D} = \{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$ of $(Y, \mathbf{X})$, the unknown parameters $\boldsymbol{\beta}(\tau)$ can be estimated by $\hat{\boldsymbol{\beta}}(\tau)$, which minimises
>
> $$R_n(\boldsymbol{\beta}, \mathcal{D}) = \sum_{i=1}^{n} \rho_\tau \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right), \tag{2.2}$$
>
> where $\rho_\tau(\mu) = \mu\{\tau - I(\mu < 0)\}^*$ is the quantile loss function given in Koenker & Bassett (1978). In the rest of the paper, we omit the τ in various expressions such as $\boldsymbol{\beta}(\tau)$ for the sake of simplicity.

- This is very weird to me, as this is done even if the likelihood is known.
- This is like always using ordinary least squares instead of GLM's.
- A very incoherent way of thinking
- The *better* way is to use the quantile function[1] which can also be justified for some discrete distr. (Padellini & R, 2018) like Poisson/Binomial/NegBinomial.

[1] *Parametric quantile regression based on the generalized gamma distribution*, by A. Noufaily, M.C. Jones, JRSS-C, 2013)

Example: Quantile regression

- Quantile regression is usually *defined* as

> Based on a random sample $\mathcal{D} = \{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$ of $(Y, \mathbf{X})$, the unknown parameters $\boldsymbol{\beta}(\tau)$ can be estimated by $\hat{\boldsymbol{\beta}}(\tau)$, which minimises
>
> $$R_n(\boldsymbol{\beta}, \mathcal{D}) = \sum_{i=1}^{n} \rho_\tau \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right), \tag{2.2}$$
>
> where $\rho_\tau(\mu) = \mu\{\tau - I(\mu < 0)\}^*$ is the quantile loss function given in Koenker & Bassett (1978). In the rest of the paper, we omit the τ in various expressions such as $\boldsymbol{\beta}(\tau)$ for the sake of simplicity.

- This is very weird to me, as this is done even if the likelihood is known.
- This is like always using ordinary least squares instead of GLM's.
- A very incoherent way of thinking
- The *better* way is to use the quantile function[1] which can also be justified for some discrete distr. (Padellini & R, 2018) like Poisson/Binomial/NegBinomial.

[1] *Parametric quantile regression based on the generalized gamma distribution*, by A. Noufaily, M.C. Jones, JRSS-C, 2013)

Example: Quantile regression

- Quantile regression is usually *defined* as

> Based on a random sample $\mathcal{D} = \{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$ of $(Y, \mathbf{X})$, the unknown parameters $\boldsymbol{\beta}(\tau)$ can be estimated by $\hat{\boldsymbol{\beta}}(\tau)$, which minimises
>
> $$R_n(\boldsymbol{\beta}, \mathcal{D}) = \sum_{i=1}^{n} \rho_\tau \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right), \tag{2.2}$$
>
> where $\rho_\tau(\mu) = \mu\{\tau - I(\mu < 0)\}^*$ is the quantile loss function given in Koenker & Bassett (1978). In the rest of the paper, we omit the τ in various expressions such as $\boldsymbol{\beta}(\tau)$ for the sake of simplicity.

- This is very weird to me, as this is done even if the likelihood is known.
- This is like always using ordinary least squares instead of GLM's.
- A very incoherent way of thinking
- The *better* way is to use the quantile function[1] which can also be justified for some discrete distr. (Padellini & R, 2018) like Poisson/Binomial/NegBinomial.

[1] *Parametric quantile regression based on the generalized gamma distribution*, by A. Noufaily, M. Jones, JRSS-C, 2013

Example: Quantile regression

- Quantile regression is usually *defined* as

> Based on a random sample $\mathcal{D} = \{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$ of $(Y, \mathbf{X})$, the unknown parameters $\boldsymbol{\beta}(\tau)$ can be estimated by $\hat{\boldsymbol{\beta}}(\tau)$, which minimises
>
> $$R_n(\boldsymbol{\beta}, \mathcal{D}) = \sum_{i=1}^{n} \rho_\tau \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right), \tag{2.2}$$
>
> where $\rho_\tau(\mu) = \mu\{\tau - I(\mu < 0)\}^*$ is the quantile loss function given in Koenker & Bassett (1978). In the rest of the paper, we omit the τ in various expressions such as $\boldsymbol{\beta}(\tau)$ for the sake of simplicity.

- This is very weird to me, as this is done even if the likelihood is known.
- This is like always using ordinary least squares instead of GLM's.
- A very incoherent way of thinking
- The *better* way is to use the quantile function[1] which can also be justified for some discrete distr. (Padellini & R, 2018) like Poisson/Binomial/NegBinomial.

[1] *Parametric quantile regression based on the generalized gamma distribution*, by A. Noufaily, M. C. Jones, JRSS-C, 2013)

Example: Quantile regression

What happen if we use the quantile-loss as the likelihood within the Bayesian framework (with scaling)?

International Statistical Review (2016), 84, 3, 327–344 doi:10.1111/insr.12114

Posterior Inference in Bayesian Quantile Regression with Asymmetric Laplace Likelihood

Yunwen Yang[1], Huixia Judy Wang[2] and Xuming He[3]

Example: Quantile regression

In short: In the limit, the mean is correct, but the variance is wrong

Based on the AL working likelihood, the posterior mean and variance of $\boldsymbol{\beta}(\tau)$ can be computed directly from the MCMC chains. Based on empirical evidence, Yu & Moyeed (2001) argued that the use of the AL likelihood is satisfactory for quantile regression, even when the likelihood is misspecified. Sriram *et al.* (2013) established sufficient conditions for the posterior consistency of model parameters in Bayesian quantile regression with the AL likelihood. However, the posterior consistency results do not imply that the interval estimates constructed from the posterior are automatically valid. It is tempting to construct interval estimates, whether they are called credible intervals or confidence intervals, from the quantiles of the posterior or by normal approximations using the variance–covariance matrix of the posterior sequence, as reported in Yu & Moyeed (2001), Li *et al.* (2010), Alhamzawi *et al.* (2012), Yue & Hong (2014) and Lum & Gelfand (2012), among others. Here, we argue that the posterior variance–covariance must be adjusted for the interval estimates to be asymptotically valid. We will present the proposed

Generalized Variational Inference

Jeremias Knoblauch
The Alan Turing Institute
Dept. of Statistics
University of Warwick
j.knoblauch@warwick.ac.uk

Jack Jewson
The Alan Turing Institute
Dept. of Statistics
University of Warwick
j.e.jewson@warwick.ac.uk

Theodoros Damoulas
The Alan Turing Institute
Depts. of Computer Science & Statistics
University of Warwick
t.damoulas@warwick.ac.uk

21 May 2019

prior belief π over $\boldsymbol{\theta}$ prevents overfitting and induces uncertainty about the optimum. This division of labour between ℓ_n and π is clearest in Zellner [87], where it is shown that q^* solves

$$\arg\min_{q\in\mathcal{P}(\boldsymbol{\Theta})}\left\{\mathbb{E}_{q(\boldsymbol{\theta})}\left[-\sum_{i=1}^{n}\log(p(x_i|\boldsymbol{\theta}))\right]+\mathrm{KLD}\left(q||\pi\right)\right\},\quad \mathrm{KLD}(q||\pi)=\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right],\quad (1)$$

for $\mathcal{P}(\boldsymbol{\Theta})$ the set of all probability distributions on $\boldsymbol{\Theta}$ and KLD the Kullback-Leibler divergence. In

We currently use this in (the next-generation) R-INLA, to

- correct the (marginal) mean
- correct the (marginal) variance (in progress)

These are highly accurate *corrections*, like essentially the same mean estimate as Laplace approximations with, at least, $\mathcal{O}(n)$ speedup.

prior belief π over $\boldsymbol{\theta}$ prevents overfitting and induces uncertainty about the optimum. This division of labour between ℓ_n and π is clearest in Zellner [87], where it is shown that q^* solves

$$\arg\min_{q\in\mathcal{P}(\boldsymbol{\Theta})}\left\{\mathbb{E}_{q(\boldsymbol{\theta})}\left[-\sum_{i=1}^{n}\log(p(x_i|\boldsymbol{\theta}))\right]+\mathrm{KLD}\left(q||\pi\right)\right\},\quad \mathrm{KLD}(q||\pi)=\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right],\quad (1)$$

for $\mathcal{P}(\boldsymbol{\Theta})$ the set of all probability distributions on $\boldsymbol{\Theta}$ and KLD the Kullback-Leibler divergence. In

We currently use this in (the next-generation) R-INLA, to

- correct the (marginal) mean
- correct the (marginal) variance (in progress)

These are highly accurate *corrections*, like essentially the same mean estimate as Laplace approximations with, at least, $\mathcal{O}(n)$ speedup.

Method	$\ell(\boldsymbol{\theta}, x_i)$	D	Π
Standard Bayes	$-\log(p(\boldsymbol{\theta}\|x_i))$	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Generalized Bayes[1]	any ℓ	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Power Bayes[2]	$-\log(p(\boldsymbol{\theta}\|x_i))$	$\frac{1}{w}$KLD, $w < 1$	$\mathcal{P}(\boldsymbol{\Theta})$
Divergence Bayes[3]	divergence-based ℓ	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Standard VI	$-\log(p(\boldsymbol{\theta}\|x_i))$	KLD	$\mathcal{Q}$
Power VI[4]	$-\log(p(\boldsymbol{\theta}\|x_i))$	$\frac{1}{w}$KLD, $w < 1$	$\mathcal{Q}$
Regularized Bayes[5]	$-\log(p(\boldsymbol{\theta}\|x_i)) + \phi(\boldsymbol{\theta}, x_i)$	KLD	$\mathcal{Q}$
(β-)VAE [6]	$-\log\left(p_{\boldsymbol{\zeta}}(x_i\|\boldsymbol{\theta})\right)$	$\beta \cdot$ KLD, $\beta > 1$	$\mathcal{Q}$
Gibbs VI[7]	any ℓ	KLD	$\mathcal{Q}$
Generalized VI	any ℓ	any D	$\mathcal{Q}$

Table 1: $P(\ell_n, D, Q)$ and relation to some existing methods. All losses have the form $\ell_n(\boldsymbol{\theta}, \boldsymbol{x}) = \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)$ for some $\ell(\boldsymbol{\theta}, x_i)$. [1][12], [2][e.g. 34, 28, 57], [3][e.g. 35, 24, 22, 40], [4][e.g. 86, 36] [5][23], but only if the regularizer can be written as $\mathbb{E}_{q(\boldsymbol{\theta})}\left[\phi(\boldsymbol{\theta}, \boldsymbol{x})\right]$ as in [88], [6][44, 32], [7][e.g. 2, 22]

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As
some unprocessed data that should have been added to the fina
extra page has been added to receive it.
If you rerun the document (without altering it) this surplus page
away, because LaTeX now knows how many pages to expect for thi
document.