

Robust non-Gaussian models and how to fit them in Stan

Rafael Cabral

rafael.medeiroscabral@kaust.edu.sa



CEMSE Department (KAUST)
Advisors: Profs. Haavard Rue and David Bolin

October 31, 2022

Paper: Controlling the flexibility of non-Gaussian processes through shrinkage priors (Cabral, Bolin, and Rue 2022)

- ▶ Interpretable parameterization of the GH distribution
- ▶ Generic class of non-Gaussian models
- ▶ Stan implementation
- ▶ Controlling flexibility with shrinkage priors

Bookdown: [Fitting robust non-Gaussian models in Stan](#)

Vignette: [Fitting robust non-Gaussian models in Stan](#)

Why go beyond Gaussian processes?

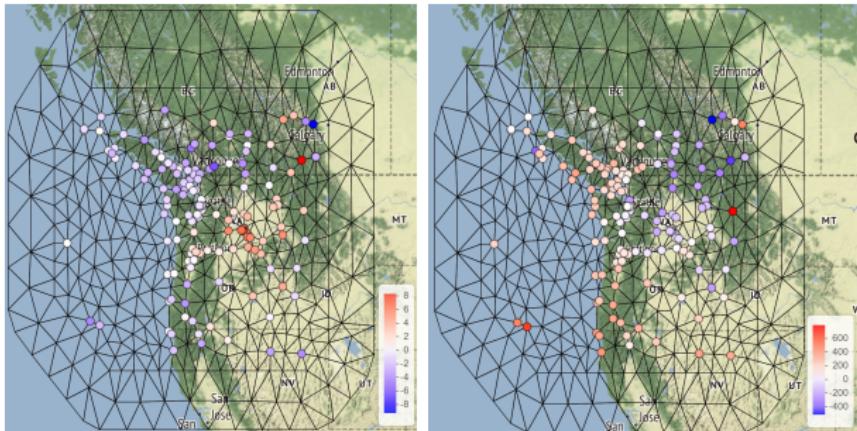


Figure 2: Measurements of temperature (left) and pressure (right)

- ▶ Gaussian processes can over-smooth in the presence of local spikes and sudden jumps in the data (Walder and Hanks 2020)
- ▶ Accommodate possible outliers in the data, and reduce their impact on the inferences (West 1984)

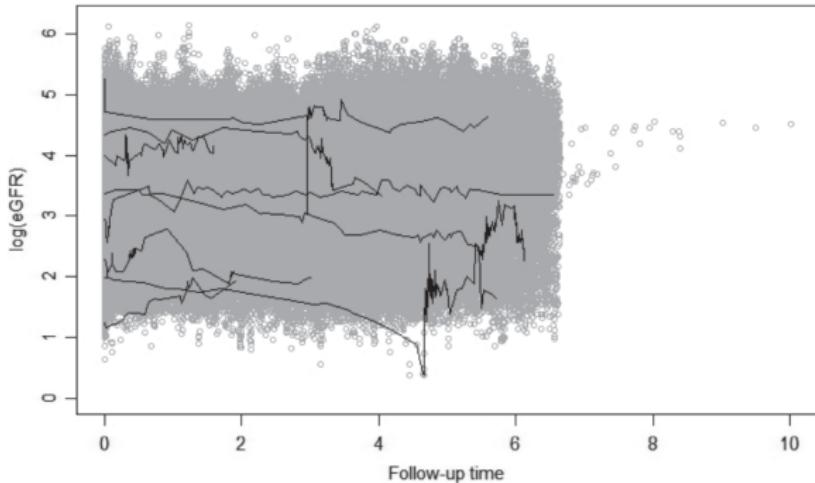


Figure 3: Log of eGFR-measurements (proxy for renal function)

Progression towards end stage renal failure (Asar et al. 2020):

- ▶ Quantity of clinical interest: Probability that renal function decreases 5% a year
- ▶ Action: Referral to specialist secondary care

If $x^G \sim N(0, Q^{-1})$, with $Q = D^T D$ it can be expressed through:

$$Dx^G \stackrel{d}{=} Z,$$

where Z is a vector of i.i.d. standard Gaussian RVs.

The non-Gaussian extension is:

$$Dx \stackrel{d}{=} \Lambda(\eta, \zeta),$$

where Λ is a vector of i.i.d. normal-inverse Gaussian (NIG) RVs.

- ▶ Mean and covariance structure are preserved
- ▶ η and ζ control the long-tailedness and skewness.
- ▶ $\eta = \zeta = 0$ leads to the Gaussian model

Example: In an RW1 process we assume $x_{i+1} - x_i \sim Z_i$, so

$$D = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & -1 & 1 & \\ & & & \ddots & \ddots \\ & & & & -1 & 1 \end{pmatrix},$$

which leads to the system $Dx = Z$. More examples:

- ▶ i.i.d. random effects
- ▶ Random walk and autoregressive processes
- ▶ SAR and CAR processes (Walder and Hanks 2020)
- ▶ Matérn processes (Wallin and Bolin 2015)
- ▶ ...

What sample paths do these models produce?

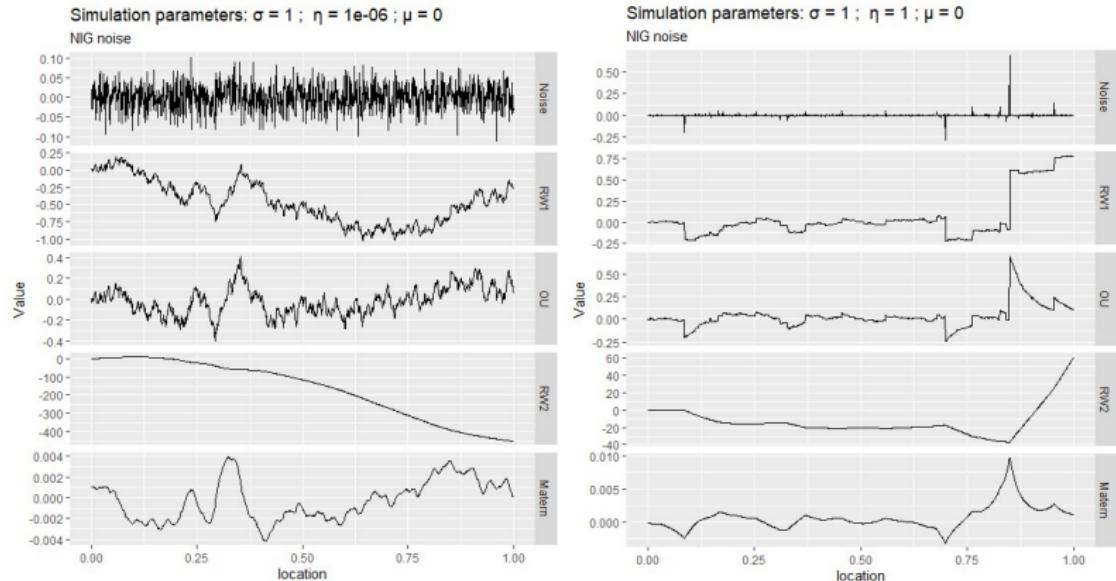


Figure 4: Sample paths of Gaussian and non-Gaussian processes.

What sample paths do these models produce?

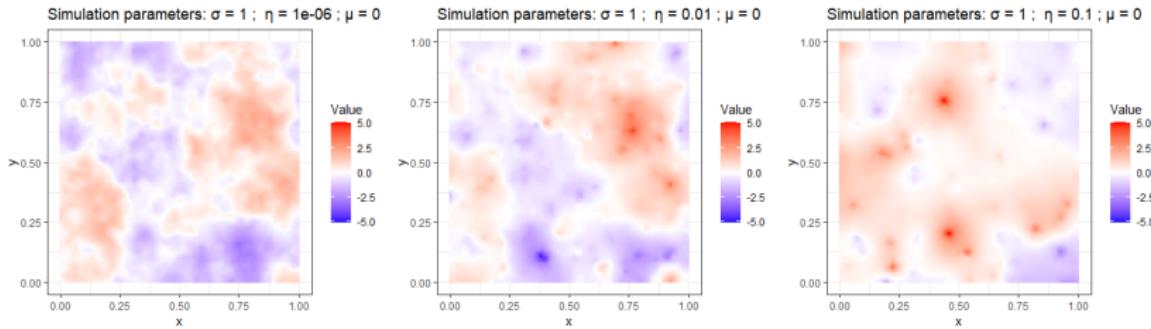


Figure 5: Sample paths of a Matérn model in 2D.

- ▶ Covariates in the mean can induce a non-stationary covariance
- ▶ The mean and covariance function of the transformed process could change in non-intuitive ways (Wallin and Bolin 2015)

With our construction, the three important model components:

- ▶ Mean
- ▶ Covariance
- ▶ Non-Gaussianity

can be modeled separately without confounding. Changing one component does not affect the others.

Declaring Gaussian ($Dx = Z$) an non-Gaussian models ($Dx = \Lambda$):

```
x ~ multi_normal_prec(rep_vector(0,N), D'*D)
```



```
x ~ nig_model(D, eta, zeta, h, 1)
```

Implementation performance:

- ▶ First attempt: 10 hours with poor diagnostics
- ▶ Right now: 15 minutes with very good diagnostics

Other implementations relied on the normal-variance mixture:

$$\begin{aligned} \mathbf{x} | \boldsymbol{\nu} &\sim N(0, Q^{-1}), \quad Q = D^T \text{diag}(\boldsymbol{\nu})^{-1} D \\ V_i &\stackrel{i.i.d.}{\sim} \text{IG}(1, \eta^{-1}) \end{aligned}$$

We can, however, "integrate out" the V 's:

$$\pi(\mathbf{x}) = |D| \prod_{i=1}^n \pi_{\Lambda_i}([Dx]_i),$$

- ▶ $\pi_{\Lambda_i}(x)$ is the pdf of a 1D NIG distribution
- ▶ We reduce the dimension by half
- ▶ Avoid inverting D

$$\log \pi(\mathbf{x}) = \log |\mathbf{D}| + \sum_{i=1}^n \log \pi_{\Lambda_i}([\mathbf{D}\mathbf{x}]_i),$$

```
real NIG_var_correction_lpdf(real x, real eta, real zeta, real h){  
    real sigmas      = 1/sqrt(1+zeta^2*eta); //variance correction  
    real hyp_alpha   = sqrt(1/eta+zeta^2)/sigmas;  
    real hyp_beta    = zeta/sigmas;  
    real hyp_delta   = sigmas*sqrt(1/eta)*h;  
    real hyp_mu      = -sigmas*zeta*h;  
    return (sqrt(hyp_alpha^2 - hyp_beta^2)*hyp_delta + hyp_beta*(x - hyp_mu) + log(hyp_alpha) +  
           log(hyp_delta) - log(pi()) - 0.5*log(hyp_delta^2 + (x - hyp_mu)^2) +  
           log(modified_bessel_second_kind(1, hyp_alpha*sqrt(hyp_delta^2 + (x - hyp_mu)^2))));  
}
```

- ▶ $\log \pi_{\Lambda_i}$ is relatively expensive to compute
- ▶ We can evaluate $\log \pi_{\Lambda_i}([\mathbf{D}\mathbf{x}]_i)$ in separate cores
- ▶ We used Stan's `reduce_sum` function

$$\log \pi(x) = \log |D| + \sum_{i=1}^n \log \pi_{\Lambda_i}([Dx]_i),$$

- Sparse matrix computations: `csr_matrix_times_vector`

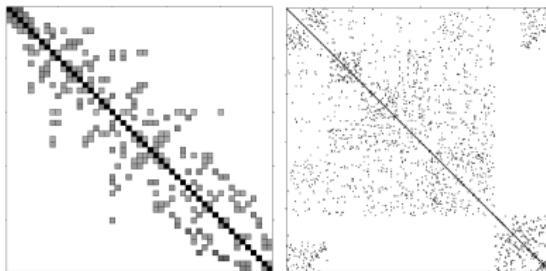
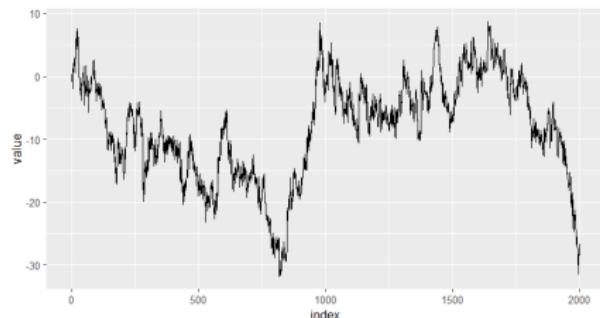


Figure 6: Matrices D for the SAR and SPDE application.

- Pre-compute your log-determinants!
 - ▶ 60% of the sampling time was spent computing determinants
 - ▶ If $D_{SAR} = I - \rho W$, then $\log |D_{SAR}| = \sum_i \log(1 - \rho v_i)$,
 $v_i = \text{eigen}(W)$

Model: Hierarchical model with a latent RW1 component.



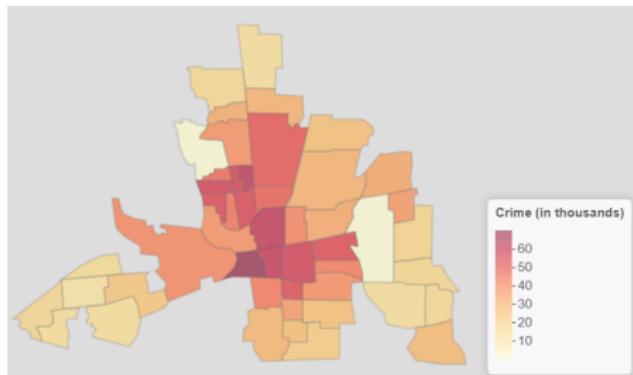
Results:

N	Implementation	ESS of η (Bulk)
2000	Variance-Mixture	0.10
	Collapsed	1.41
	Collapsed + Sparse + Parallel	5.30
4000	Variance-Mixture	0.01
	Collapsed	0.32
	Collapsed + Sparse + Parallel	2.25

Model:

$$y_i = \beta_0 + \beta_1 \text{HV}_i + \beta_2 \text{HI}_i + \sigma x_i,$$

- ▶ y_i : crime rates in thousands in 49 counties of Columbus, Ohio
- ▶ HV_i, HI_i : Household value and income
- ▶ x is a spatial effects SAR model ($D_{SAR} = I - \rho W$)



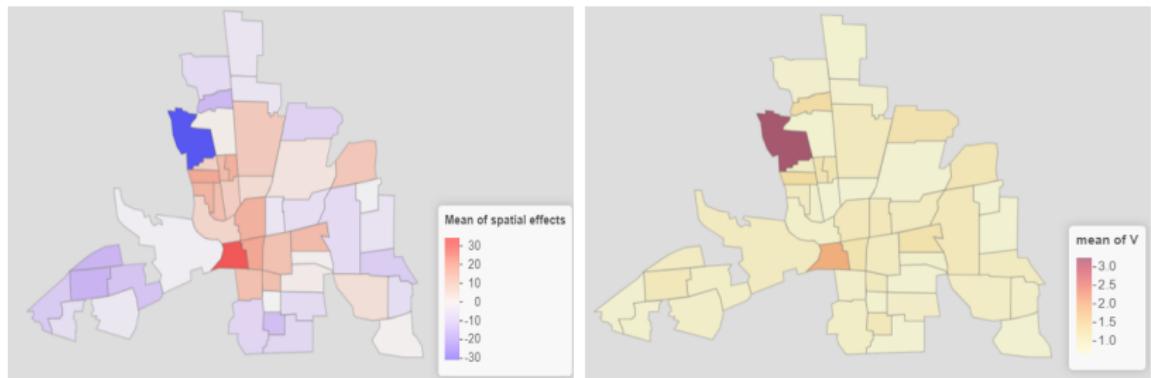
Model: $y = B\beta + \sigma x$

- ▶ B = design matrix
 - ▶ $D_{SAR}x = \Lambda$, $D_{SAR} = I - \rho W$
-

```
transformed parameters{  
    vector[N] x = (y - B*beta)/sigma;      // Spatial effects  
}  
  
model{  
    matrix[N,N] D = add_diag(-rho*W, 1); // D = I - rho W  
    x ~ multi_normal_prec(rep_vector(0,N), D'*D); // Gaussian  
    x ~ nig_model(D, eta, zeta, h, 1);           // NIG  
    ...  
}
```

Application 1: SAR model

We can identify the “outlier counties” through $E[V|y]$.



- Setup: 4 parallel chains; 2000 warmup, 3000 sampling iterations
- Gaussian model: 34s; NIG model: 93s (Collapsed only)

variable	σ	ρ	η	ζ	β_0	HV	HI
Summary	27.41	0.51	6.51	-0.11	59.30	-0.15	-1.45
q05	22.10	0.15	2.95	-1.53	33.02	-0.39	-2.38
q95	33.10	0.80	10.94	1.27	81.45	0.07	-0.50
rhat	1	1	1	1	1	1	1
ess_bulk	4578	3505	5081	1895	2091	3542	2892

$X(t)$ is a Matérn Gaussian process (Whittle 1963):

$$(\kappa^2 - \Delta)^{\alpha/2} X(s) = \sigma \mathcal{W}(s), \quad s \in \mathbb{R}^d,$$

- ▶ κ is a spatial range parameter, α is a smoothness parameter
- ▶ $\Delta = \sum_i \partial^2 / \partial x_i^2$ is the Laplace operator
- ▶ $\mathcal{W}(s)$ is a Gaussian white noise process
- Non-Gaussian extension: $\mathcal{W}(s)$ becomes $\Lambda(s)$
- Approximation to discrete space (Bolin 2014): $Dx = \Lambda$ (FEM)

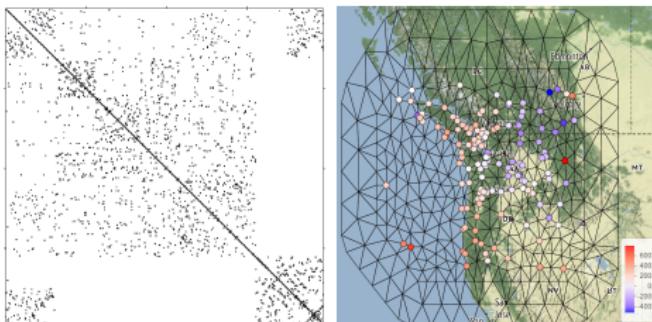


Figure 7: Matrix D (right), temperature data and mesh triangulation (right).

Application 2: Spatial Matérn model

Model: $y_i = x_i + \sigma_\epsilon \epsilon_i$, x is a Mátern non-Gaussian spatial field.

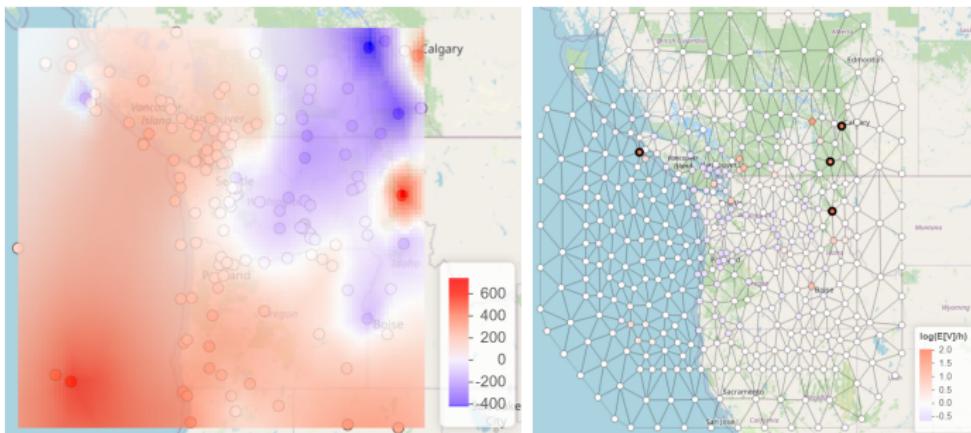


Figure 8: Posterior mean of x (left) and V (right)

Model	MSE	MAE	CRPS
Gaussian model	24036.848	98.647	54.893
NIG model	20246.371	87.207	53.339

Table 1: Mean squared error (MSE), mean absolute error (MAE), and continuously ranked probability score (CRPS) for the leave-one-out cross-validation predictions.

LnGMS: Latent non-Gaussian Models

Response	$y x, \theta_1 \sim \prod_{i \in \mathcal{I}} \pi(y_i x_i, \theta_1)$
Latent field	$x \mathbf{V}, \theta_2 \sim N(0, D(\theta_2)^{-1} \text{diag}(\mathbf{V}) D(\theta_2)^{-T})$
Mixing variables	$V_i \eta \stackrel{\text{ind.}}{\sim} \text{IG}(1, \eta^{-1})$
Hyperparameters	$\theta \sim \pi(\theta), \quad \eta \sim \text{Exp}(\alpha_\eta)$

- ▶ We find: $q(x, \theta, V) = q(x, \theta)q(V)$
- ▶ that minimizes: $\text{KLD}(q(x, \theta, V) | \pi(x, \theta, V|y))$
- ▶ We get an algorithm that recursively fits LGMs until convergence
- ▶ To fit each LGM we use INLA (Rue et al. 2017)

- Implemented in the `ngvb` package (coming soon!).
- Example: Fitting a latent AR1 process.

```
data      <- list(x = 1:100, y = g(1:100))
formula <- y ~ f(x, model = "ar1")
LGM      <- inla(formula, data = data)
LnGM     <- ngvb(fit = LGM, selection = list(x=1:100))
```

Berger 2013:

It behooves Bayesians to provide simple “standardized” Bayesian procedures with built-in robustness. While it is certainly desirable to perform specific robustness studies (...) such will require a fairly high level of sophistication. For routine use by non-experts, Bayesian procedures are needed which require minimal prior inputs and are, in some sense, inherently robust.

Thank you for your attention!

- Asar, Özgür et al. (2020). "Linear mixed effects models for non-Gaussian continuous repeated measurement data". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 69.5, pp. 1015–1065.
- Berger, James O (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.
- Bolin, David (2014). "Spatial Matérn fields driven by non-Gaussian noise". In: *Scandinavian Journal of Statistics* 41.3, pp. 557–579.
- Cabral, Rafael, David Bolin, and Håvard Rue (2022). "Controlling the flexibility of non-Gaussian processes through shrinkage priors". In: *arXiv preprint arXiv:2203.05510*.
- Rue, Håvard et al. (2017). "Bayesian computing with INLA: a review". In: *Annual Review of Statistics and Its Application* 4, pp. 395–421.
- Walder, Adam and Ephraim M Hanks (2020). "Bayesian analysis of spatial generalized linear mixed models with Laplace moving average random fields". In: *Computational Statistics & Data Analysis* 144, p. 106861.

- Wallin, Jonas and David Bolin (2015). "Geostatistical modelling using non-Gaussian Matérn fields". In: *Scandinavian Journal of Statistics* 42.3, pp. 872–890.
- West, Mike (1984). "Outlier models and prior distributions in Bayesian linear regression". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 46.3, pp. 431–439.
- Whittle, Peter (1963). "Stochastic processes in several dimensions". In: *Bulletin of the International Statistical Institute* 40.2, pp. 974–994.