

Web-based Supporting Material for *Robust Fitting and Smoothing Parameter Selection for Generalized Additive Models for Location, Scale and Shape*

William H. Aeberhard¹, Eva Cantoni², Giampiero Marra³, and Rosalba Radice⁴

¹Swiss Data Science Center, ETH Zurich

²Research Center for Statistics and GSEM, University of Geneva

³Department of Statistical Science, University College London

⁴Cass Business School, City, University of London

Web Appendix A: Conditions and Proof of Theorem 1

The following conditions are required for Theorem 1 in the main body of the paper.

- (C1) For any $c < \infty$, $\rho_c : \mathbb{R} \rightarrow \mathbb{R}$ is convex, monotonically increasing, twice continuously differentiable, and has a first derivative ρ'_c that is bounded within $[0, 1]$.
- (C2) The parameter space $\Delta \subset \mathbb{R}^p$ is compact and δ_0 is an interior point of Δ .
- (C3) For any y in the support, the mapping $\delta \rightarrow f(y|\mu, \sigma, \nu)$ is twice continuously differentiable with respect to δ and $\sup_{\delta \in \Delta} f(y|\mu, \sigma, \nu) < \infty$.
- (C4) For all δ in some open neighborhood around δ_0 , $\mathbf{M}(\delta)$ is positive definite and $\mathbf{Q}(\delta) < \infty$.
- (C5) The smoothness parameter $\lambda = o(n^{1/2})\mathbf{1}$, i.e. a vanishing penalty $n^{-1}\lambda = o(n^{-1/2})\mathbf{1}$ for the log-likelihood scaled by $1/n$.

Condition (C1) summarizes the requirements for the user-specified ρ function. In particular, the boundedness of the first derivative is the key to the robustness property of the proposed estimator as it ensures that the influence function is itself bounded (see below).

These requirements are satisfied by the log-logistic function used in the main body of the paper.

The conditions in (C2) essentially guarantee that the asymptotic distribution is absolutely continuous (i.e. without point-masses on the boundary of the domain). The compactness requirement is rather standard in non-parametric and sieve estimation contexts.

Condition (C3) includes most commonly used families of distributions. We note it may be relaxed to Lipschitz requirements on the derivative of $\log f(y|\mu, \sigma, \nu)$ and may hold only for asymptotically non-null sets (see Huber, 1967).

The requirements in (C4) are standard minimal conditions in the robustness literature, akin to assuming that the Fisher information matrix exists in a neighborhood of the true parameter in a maximum likelihood (ML) framework.

Finally, (C5) imposes a minimum rate at which $\boldsymbol{\lambda}$ decreases as $n \rightarrow \infty$. This rate is necessary to guarantee that the penalization does not induce any asymptotic bias for the robust estimator.

Proof of Theorem 1. The penalized robust estimator $\hat{\boldsymbol{\delta}}$ is the solution in $\boldsymbol{\delta}$ to

$$\mathbf{0} = \frac{\partial \tilde{\ell}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} - \frac{1}{n} \mathbf{S}(\boldsymbol{\lambda}) \boldsymbol{\delta}, \quad (1)$$

where here $\tilde{\ell}(\boldsymbol{\delta}) = \frac{1}{n} \sum_{i=1}^n \rho_c(\ell(\boldsymbol{\delta})_i) - \frac{1}{n} b_\rho(\boldsymbol{\delta})$ is the scaled robustified log-likelihood. Given the smoothness of $\tilde{\ell}(\boldsymbol{\delta})$ ensured by (C1) and (C3), a first-order Taylor expansion about $\boldsymbol{\delta}_0$ of the right-hand side of (1) evaluated at the solution $\hat{\boldsymbol{\delta}}$ yields

$$\mathbf{0} = \left. \frac{\partial \tilde{\ell}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \right|_{\boldsymbol{\delta}=\boldsymbol{\delta}_0} - n^{-1} \mathbf{S}(\boldsymbol{\lambda}) \boldsymbol{\delta}_0 + \left(\left. \frac{\partial^2 \tilde{\ell}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^\top} \right|_{\boldsymbol{\delta}=\boldsymbol{\delta}_0} - n^{-1} \mathbf{S}(\boldsymbol{\lambda}) \right) (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) + O_p(\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0\|_2^2).$$

Multiplying by $\sqrt{n}$ on both sides and rearranging lead to

$$\begin{aligned} \left(\frac{\partial^2 \tilde{\ell}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^\top} \bigg|_{\boldsymbol{\delta}=\boldsymbol{\delta}_0} - n^{-1} \mathbf{S}(\boldsymbol{\lambda}) \right) \sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \\ = -\sqrt{n} \frac{\partial \tilde{\ell}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \bigg|_{\boldsymbol{\delta}=\boldsymbol{\delta}_0} + n^{-1/2} \mathbf{S}(\boldsymbol{\lambda}) \boldsymbol{\delta}_0 + O_p(\sqrt{n} \|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0\|_2^2). \end{aligned} \quad (2)$$

Considering the limit as $n \rightarrow \infty$, (C5) implies $n^{-1} \mathbf{S}(\boldsymbol{\lambda}) \rightarrow \mathbf{0}$ and $n^{-1/2} \mathbf{S}(\boldsymbol{\lambda}) \boldsymbol{\delta}_0 \rightarrow \mathbf{0}$. This effectively reduces the estimating equations to those of the unpenalized robust estimator that solves $\partial \tilde{\ell}(\boldsymbol{\delta}) / \partial \boldsymbol{\delta} = \mathbf{0}$. Viewing $\boldsymbol{\delta}_0$ as the true parameter that generated the data, we can thus invoke standard results from M -estimation theory. In particular, conditions (C1)–(C4) are sufficient to apply Theorem 2 of Huber (1967) so that the unpenalized robust estimator converges a.s. to $\boldsymbol{\delta}_0$. This implies that the statistical functional $\mathbf{T}(F)$ defined as the solution in $\boldsymbol{\delta}$ to $\mathbb{E}_F [\psi(Y, \boldsymbol{\delta})] = \mathbf{0}$, where

$$\psi(Y, \boldsymbol{\delta}) = \rho'_c(\log f(Y|\mu, \sigma, \nu)) \frac{\partial \log f(Y|\mu, \sigma, \nu)}{\partial \boldsymbol{\delta}} - \frac{\partial}{\partial \boldsymbol{\delta}} \mathbb{E} [\rho_c^*(\log f(Y|\mu, \sigma, \nu))],$$

satisfies $\mathbf{T}(D) = \boldsymbol{\delta}_0$, i.e. it is Fisher consistent for $\boldsymbol{\delta}_0$. Asymptotic normality follows under (C1)–(C4) by observing that the remainder in (2) is $O_p(n^{-1/2})$ because of $\sqrt{n}$ -consistency and by Theorem 3 of Huber (1967):

$$\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \xrightarrow[n \rightarrow \infty]{D} \mathbf{N}(\mathbf{0}, \mathbf{V}(\boldsymbol{\delta}_0))$$

where $\mathbf{V}(\boldsymbol{\delta}) = \mathbf{M}(\boldsymbol{\delta})^{-1} \mathbf{Q}(\boldsymbol{\delta}) \mathbf{M}(\boldsymbol{\delta})^{-\top}$. □

Web Appendix B: Extended Fellner-Schall Method

We follow here the main steps of Wood and Fasiolo (2017) in developing a robust extended Fellner-Schall (EFS) method to select the smoothness parameter $\boldsymbol{\lambda}$.

The quadratic penalty in the penalized robustified log-likelihood

$$\tilde{\ell}_p(\boldsymbol{\delta}) = \tilde{\ell}(\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\delta}^\top \mathbf{S}\boldsymbol{\delta}$$

can be viewed as an improper Gaussian prior distribution on $\boldsymbol{\delta}$, seen as a random vector throughout this section, with $\boldsymbol{\lambda}$ now seen as another model parameter. For this, consider adding the term corresponding to the determinant of the Gaussian covariance matrix to get an improper prior, i.e. up to a constant term (involving $\sqrt{2\pi}$). We thus define the joint robustified log-likelihood l for $\mathbf{Y} = (Y_1, \dots, Y_n)$ and $\boldsymbol{\delta}$ as

$$l(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\lambda}) = \tilde{\ell}_p(\boldsymbol{\delta}) - \frac{1}{2} \log |\mathbf{S}(\boldsymbol{\lambda})^{-1}| = \tilde{\ell}(\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\delta}^\top \mathbf{S}(\boldsymbol{\lambda})\boldsymbol{\delta} + \frac{1}{2} \log |\mathbf{S}(\boldsymbol{\lambda})|$$

where $|\cdot|$ denotes matrix determinant, so that we can view $L(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\lambda}) = \exp(l(\mathbf{y}, \boldsymbol{\delta}))$ as a joint (robustified) likelihood up to constant terms:

$$L(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\lambda}) = \exp(\tilde{\ell}(\boldsymbol{\delta})) \times \exp\left(-\frac{1}{2}\boldsymbol{\delta}^\top \mathbf{S}(\boldsymbol{\lambda})\boldsymbol{\delta}\right) \frac{1}{\sqrt{|\mathbf{S}(\boldsymbol{\lambda})^{-1}|}}. \quad (3)$$

The multivariate Gaussian distribution of $\boldsymbol{\delta}$ has a mean vector of zero and covariance $\mathbf{S}(\boldsymbol{\lambda})^{-1}$, or precision matrix $\mathbf{S}(\boldsymbol{\lambda})$. By construction, $\mathbf{S}(\boldsymbol{\lambda})$ is symmetric and positive definite so that all its eigenvalues are strictly positive and it is thus invertible. Note that this added term does not depend on $\boldsymbol{\delta}$ so that our robust estimator $\hat{\boldsymbol{\delta}}$ can also be defined as $\arg \max_{\boldsymbol{\delta}} l(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\lambda})$.

We derive the marginal robustified likelihood for $\mathbf{Y}$ by integrating out $\boldsymbol{\delta}$ from $L(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\lambda})$: $L(\mathbf{y}; \boldsymbol{\lambda}) = \int_{\Delta} L(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\lambda}) d\boldsymbol{\delta}$. The Laplace approximation of this integral can be summarized as a second-order Taylor expansion of $l(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\lambda})$ about the global maximum $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}}(\boldsymbol{\lambda}) = \arg \max_{\boldsymbol{\delta}} l(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\lambda})$ for a given $\boldsymbol{\lambda}$ and observation vector $\mathbf{y}$. This yields the following approximation to the marginal robustified log-likelihood

$$\log L(\mathbf{y}; \boldsymbol{\lambda}) \approx l_{\text{LA}}(\mathbf{y}; \boldsymbol{\lambda}) = \tilde{\ell}(\hat{\boldsymbol{\delta}}) - \frac{1}{2}\hat{\boldsymbol{\delta}}^\top \mathbf{S}(\boldsymbol{\lambda})\hat{\boldsymbol{\delta}} + \frac{1}{2} \log |\mathbf{S}(\boldsymbol{\lambda})| - \frac{1}{2} \log |\widehat{\mathbf{M}}_p(\hat{\boldsymbol{\delta}}, \boldsymbol{\lambda})|, \quad (4)$$

where the “LA” subscript identifies Laplace-approximated quantities and the negative Hes-

sian

$$\widehat{\mathbf{M}}_p(\boldsymbol{\delta}, \boldsymbol{\lambda}) = \frac{\partial^2 l(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\lambda})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^\top} = -\frac{\partial^2 \tilde{\ell}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^\top} + \mathbf{S}(\boldsymbol{\lambda}) = \widehat{\mathbf{M}}(\boldsymbol{\delta}) + \mathbf{S}(\boldsymbol{\lambda})$$

is guaranteed to be positive definite at least in a neighborhood of $\hat{\boldsymbol{\delta}}$ for sufficiently large n under the conditions of Theorem 1.

Now, viewing $\boldsymbol{\lambda}$ as the parameter of this (Laplace-approximated) marginal log-likelihood, ideally we would then compute the first-order derivative:

$$\frac{\partial l_{\text{LA}}(\mathbf{y}; \boldsymbol{\lambda})}{\partial \lambda_j} = \frac{\partial \tilde{\ell}(\hat{\boldsymbol{\delta}}(\boldsymbol{\lambda}))}{\partial \lambda_j} - \frac{1}{2} \frac{\partial}{\partial \lambda_j} \hat{\boldsymbol{\delta}}(\boldsymbol{\lambda})^\top \mathbf{S}(\boldsymbol{\lambda}) \hat{\boldsymbol{\delta}}(\boldsymbol{\lambda}) + \frac{1}{2} \frac{\partial}{\partial \lambda_j} \log |\mathbf{S}(\boldsymbol{\lambda})| - \frac{1}{2} \frac{\partial}{\partial \lambda_j} \log |\widehat{\mathbf{M}}_p(\hat{\boldsymbol{\delta}}, \boldsymbol{\lambda})|. \quad (5)$$

Unfortunately, it cannot be computed directly in general since $\boldsymbol{\lambda}$ appears essentially everywhere. In particular, $\boldsymbol{\lambda}$ appears in $\widehat{\mathbf{M}}_p$ both explicitly because $\widehat{\mathbf{M}}_p$ is a direct function of $\mathbf{S}(\boldsymbol{\lambda})$ (see above) and implicitly through $\hat{\boldsymbol{\delta}}(\boldsymbol{\lambda})$. We therefore need to neglect the dependence on $\boldsymbol{\lambda}$ for some terms in (5). Specifically, we are going to consider $\hat{\boldsymbol{\delta}}$ as a function of a previous “inactive” iterate $\boldsymbol{\lambda}'$, so that $\hat{\boldsymbol{\delta}}$ is fixed with respect to the current “active” $\boldsymbol{\lambda}$. This implies that $\widehat{\mathbf{M}}_p(\hat{\boldsymbol{\delta}}, \boldsymbol{\lambda}) = -\left. \frac{\partial^2 \tilde{\ell}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^\top} \right|_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}} + \mathbf{S}(\boldsymbol{\lambda})$ is now a function of the active $\boldsymbol{\lambda}$ only explicitly through $\mathbf{S}(\boldsymbol{\lambda})$. That way, the new Laplace-approximated marginal log-likelihood, written as $l_{\text{LA}}(\mathbf{y}; \boldsymbol{\lambda}, \boldsymbol{\lambda}')$, is now a function of the active $\boldsymbol{\lambda}$ only through $\mathbf{S}(\boldsymbol{\lambda})$. To compute the derivative of $l_{\text{LA}}(\mathbf{y}; \boldsymbol{\lambda}, \boldsymbol{\lambda}')$ with respect to the active $\boldsymbol{\lambda}$, we adapt Jacobi’s formula to invertible matrices and obtain

$$\frac{\partial \log |\mathbf{S}(\boldsymbol{\lambda})|}{\partial \lambda_j} = \text{tr} \left\{ \mathbf{S}(\boldsymbol{\lambda})^{-1} \frac{\partial \mathbf{S}(\boldsymbol{\lambda})}{\partial \lambda_j} \right\}$$

since $\mathbf{S}(\boldsymbol{\lambda})$ is symmetric and invertible (and a differentiable function of $\boldsymbol{\lambda}$). Based on this result, we can get a simplified version of (5):

$$\frac{\partial l_{\text{LA}}(\mathbf{y}; \boldsymbol{\lambda}, \boldsymbol{\lambda}')}{\partial \lambda_j} = -\frac{1}{2} \hat{\boldsymbol{\delta}}^\top \frac{\partial \mathbf{S}(\boldsymbol{\lambda})}{\partial \lambda_j} \hat{\boldsymbol{\delta}} + \frac{1}{2} \text{tr} \left\{ \mathbf{S}(\boldsymbol{\lambda})^{-1} \frac{\partial \mathbf{S}(\boldsymbol{\lambda})}{\partial \lambda_j} \right\} - \frac{1}{2} \text{tr} \left\{ \widehat{\mathbf{M}}_p(\hat{\boldsymbol{\delta}}, \boldsymbol{\lambda})^{-1} \frac{\partial \mathbf{S}(\boldsymbol{\lambda})}{\partial \lambda_j} \right\}. \quad (6)$$

In this new expression, $\partial \mathbf{S}(\boldsymbol{\lambda}) / \partial \lambda_j$ is straightforward to write down and implement since $\mathbf{S}(\boldsymbol{\lambda})$ is block-diagonal and each block is linear in the components of $\boldsymbol{\lambda}$ and only involves the

(known) basis functions.

Now we can follow the same heuristic reasoning as in Section 2 of Wood and Fasiolo (2017). There are four necessary requirements for the update of λ_j :

1. λ_j must not change if $\partial l_{\text{LA}}(\mathbf{y}; \boldsymbol{\lambda}, \boldsymbol{\lambda}')/\partial \lambda_j = 0$.
2. λ_j needs to be decreased (because we maximize l_{LA}) if $\partial l_{\text{LA}}(\mathbf{y}; \boldsymbol{\lambda}, \boldsymbol{\lambda}')/\partial \lambda_j < 0$, this happens iff

$$\text{tr}\{\mathbf{S}(\boldsymbol{\lambda})^{-1}\partial\mathbf{S}(\boldsymbol{\lambda})/\partial\lambda_j\} - \text{tr}\{\widehat{\mathbf{M}}_p(\hat{\boldsymbol{\delta}}, \boldsymbol{\lambda})^{-1}\partial\mathbf{S}(\boldsymbol{\lambda})/\partial\lambda_j\} < \hat{\boldsymbol{\delta}}^\top (\partial\mathbf{S}(\boldsymbol{\lambda})/\partial\lambda_j)\hat{\boldsymbol{\delta}}.$$

3. Similarly, λ_j needs to be increased if $\partial l_{\text{LA}}(\mathbf{y}; \boldsymbol{\lambda}, \boldsymbol{\lambda}')/\partial \lambda_j > 0$, this happens iff

$$\text{tr}\{\mathbf{S}(\boldsymbol{\lambda})^{-1}\partial\mathbf{S}(\boldsymbol{\lambda})/\partial\lambda_j\} - \text{tr}\{\widehat{\mathbf{M}}_p(\hat{\boldsymbol{\delta}}, \boldsymbol{\lambda})^{-1}\partial\mathbf{S}(\boldsymbol{\lambda})/\partial\lambda_j\} > \hat{\boldsymbol{\delta}}^\top (\partial\mathbf{S}(\boldsymbol{\lambda})/\partial\lambda_j)\hat{\boldsymbol{\delta}}.$$

4. λ_j must remain positive.

Before moving on, we need to check four conditions to apply Theorem 1 of Wood and Fasiolo (2017): (i) the negative Hessian $-\frac{\partial^2 \tilde{\ell}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^\top}$ is positive definite; (ii) $\mathbf{S}(\boldsymbol{\lambda})$ is positive semi-definite; (iii) the null space of $\mathbf{S}(\boldsymbol{\lambda})$ is independent of $\boldsymbol{\lambda}$; (iv) $\partial\mathbf{S}(\boldsymbol{\lambda})/\partial\lambda_j$ is positive semi-definite for all j . Condition (i) is satisfied under the conditions of our Theorem 1 in the main body of the paper, at least in a neighborhood of $\hat{\boldsymbol{\delta}}$ for sufficiently large n based on some reasonable previous iterate $\boldsymbol{\lambda}'$. Condition (ii) directly follows from our construction of $\mathbf{S}(\boldsymbol{\lambda})$ as a positive definite matrix. Condition (iii) holds because $\mathbf{S}(\boldsymbol{\lambda})$ is invertible and thus its null space is $\{\mathbf{0}\}$. Finally, for Condition (iv), recall that $\mathbf{S}(\boldsymbol{\lambda})$ is typically block-diagonal and that its blocks (themselves block-diagonal) are usually linear in all elements of $\boldsymbol{\lambda}$. The coefficients typically are integrated quadratic forms of the basis functions (we penalize the wiggleness of the fitted smooth function), so the derivative $\partial\mathbf{S}(\boldsymbol{\lambda})/\partial\lambda_j$ typically is positive definite so that Condition (iv) is satisfied. Hence Theorem 1 in Wood and Fasiolo (2017) holds: $\text{tr}\{\mathbf{S}(\boldsymbol{\lambda})^{-1}\partial\mathbf{S}(\boldsymbol{\lambda})/\partial\lambda_j\} - \text{tr}\{\widehat{\mathbf{M}}_p(\hat{\boldsymbol{\delta}}, \boldsymbol{\lambda})^{-1}\partial\mathbf{S}(\boldsymbol{\lambda})/\partial\lambda_j\} > 0$. Further, by the same arguments for checking Condition (iv), the quadratic form $\hat{\boldsymbol{\delta}}^\top (\partial\mathbf{S}(\boldsymbol{\lambda})/\partial\lambda_j)\hat{\boldsymbol{\delta}}$ is always positive.

This allows us to consider an analogous expression to that of Equation (8) in Wood and Fasiolo (2017) to update all elements of $\boldsymbol{\lambda}$ from iteration $[k]$ to $[k + 1]$:

$$\lambda_j^{[k+1]} = \lambda_j^{[k]} \times \frac{\text{tr}\{\mathbf{S}(\boldsymbol{\lambda}^{[k]})^{-1} \partial \mathbf{S}(\boldsymbol{\lambda}) / \partial \lambda_j |_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[k]}}\} - \text{tr}\{\widehat{\mathbf{M}}_p(\hat{\boldsymbol{\delta}}, \boldsymbol{\lambda}^{[k]})^{-1} \partial \mathbf{S}(\boldsymbol{\lambda}) / \partial \lambda_j |_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[k]}}\}}{\hat{\boldsymbol{\delta}}^\top (\partial \mathbf{S}(\boldsymbol{\lambda}) / \partial \lambda_j |_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[k]}}) \hat{\boldsymbol{\delta}}}, \quad (7)$$

where $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}}(\boldsymbol{\lambda}^{[k]})$ here. This heuristic update satisfies all four requirements above since the ratio multiplying $\lambda_j^{[k]}$ is always positive and equals one if $\partial l_{\text{LA}}(\mathbf{y}; \boldsymbol{\lambda}, \boldsymbol{\lambda}') / \partial \lambda_j = 0$. As noted by Wood and Fasiolo (2017, p. 1073), this ratio is not guaranteed to have the optimal step size, e.g. we could under-shoot the optimal update in terms of increasing $l_{\text{LA}}(\mathbf{y}; \boldsymbol{\lambda}, \boldsymbol{\lambda}')$. So a line search at each update may be useful, although the additional computational cost might offset the gain in reducing the number of iterations.

Web Appendix C: Supplementary Figures

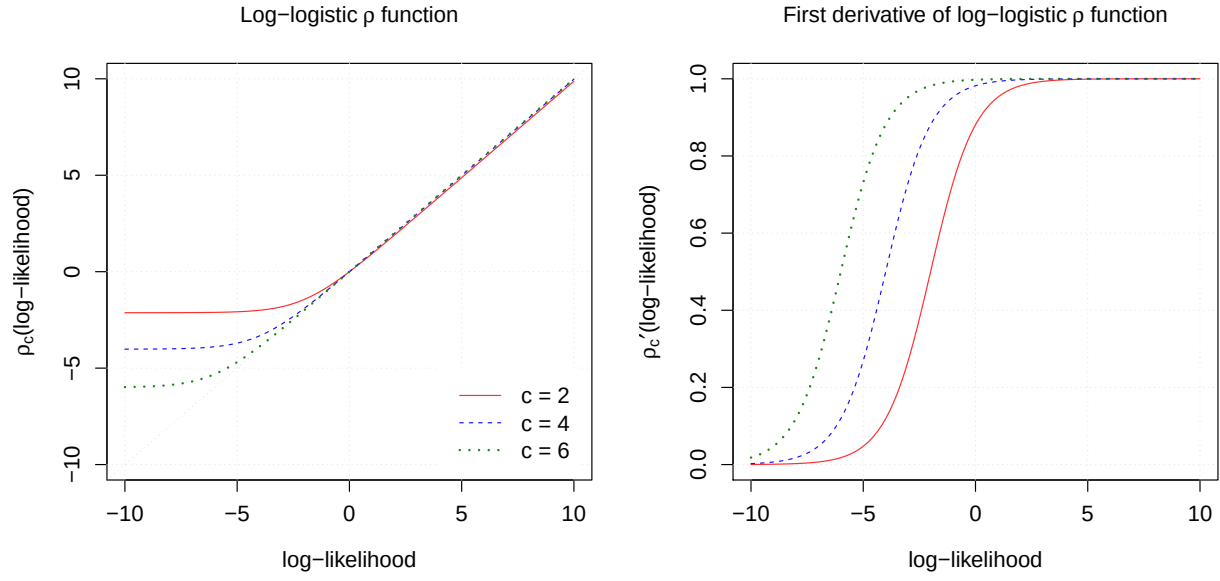


Figure S1: Log-logistic ρ_c function (left panel) and its first derivative ρ'_c (right panel), for some values of the robustness tuning constant c .

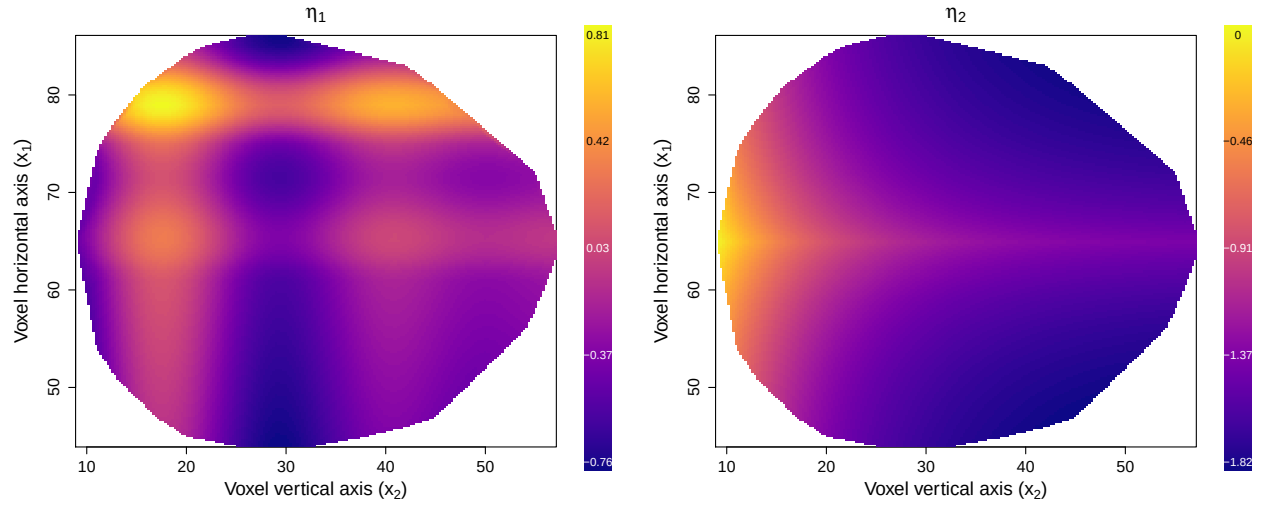


Figure S2: GAMLSS simulation, surfaces of s_1 and s_2 functions used to simulate data.

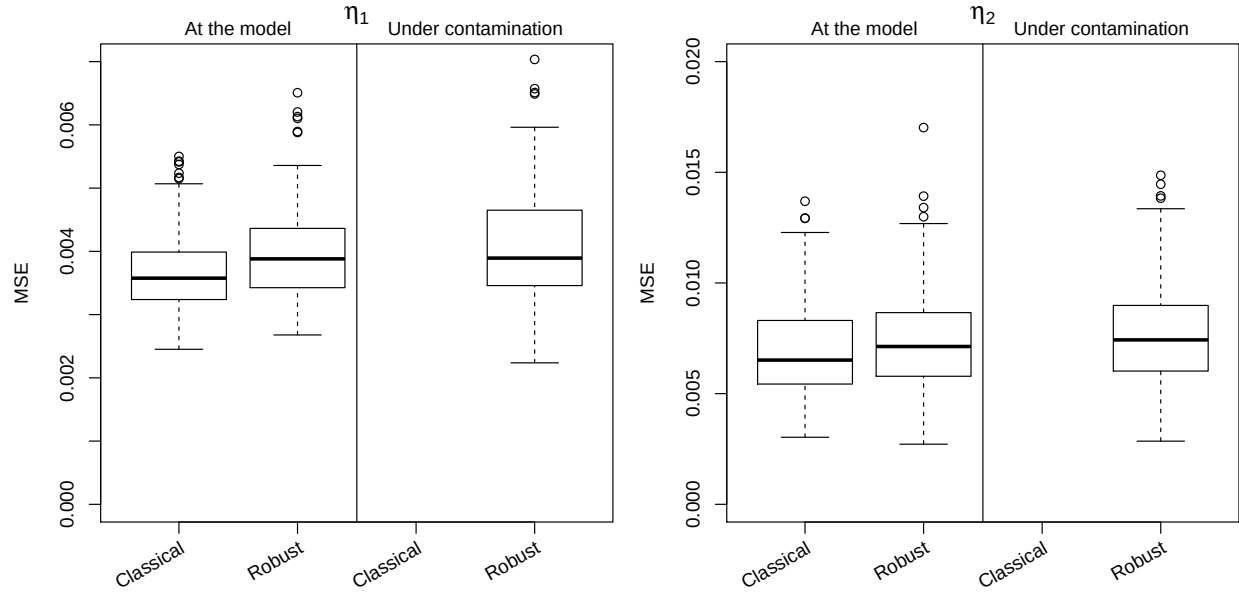


Figure S3: GAMLSS simulation, MSE for the linear predictors η_1 (left panel) and η_2 (right panel) for classical and robust methods with data generated at the assumed model and under contamination. Vertical scales set manually for better visualization.

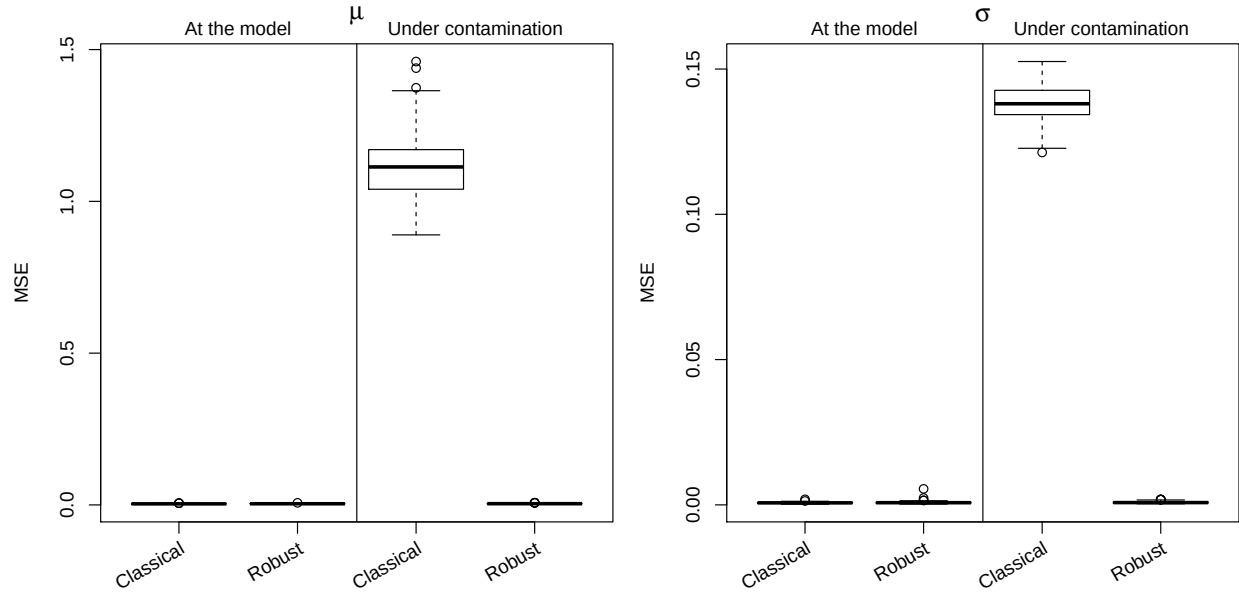


Figure S4: GAMLSS simulation, MSE for the canonical parameters μ (left panel) and σ (right panel) for classical and robust methods with data generated at the assumed model and under contamination.

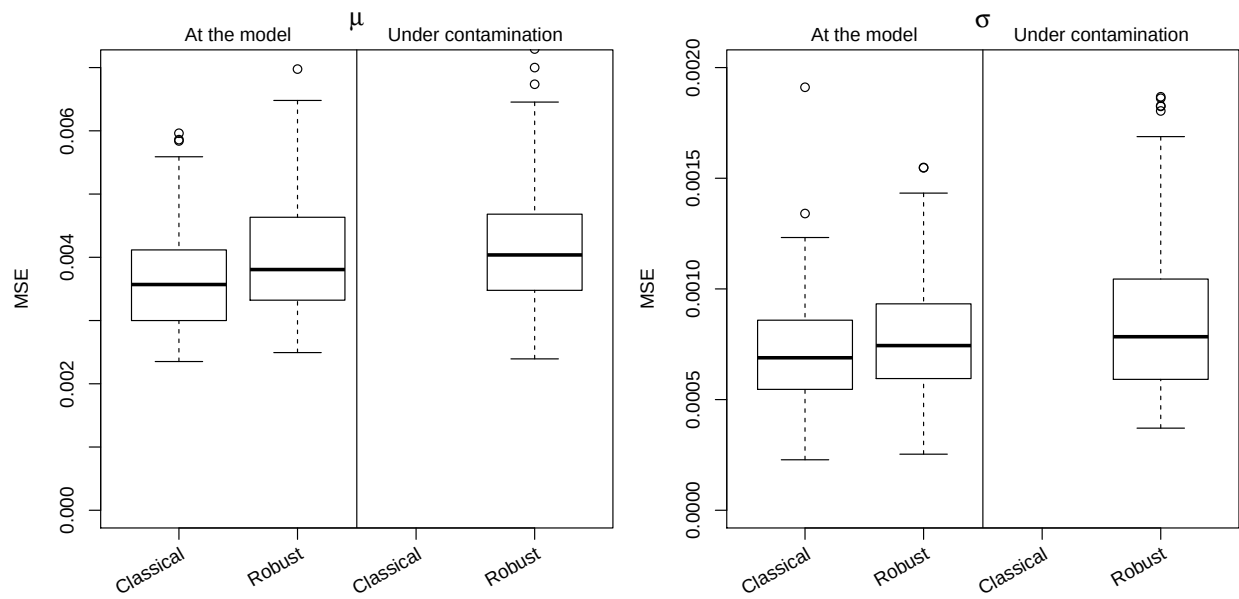


Figure S5: GAMLSS simulation, MSE for the canonical parameters μ (left panel) and σ (right panel) for classical and robust methods with data generated at the assumed model and under contamination. Vertical scales set manually for better visualization.

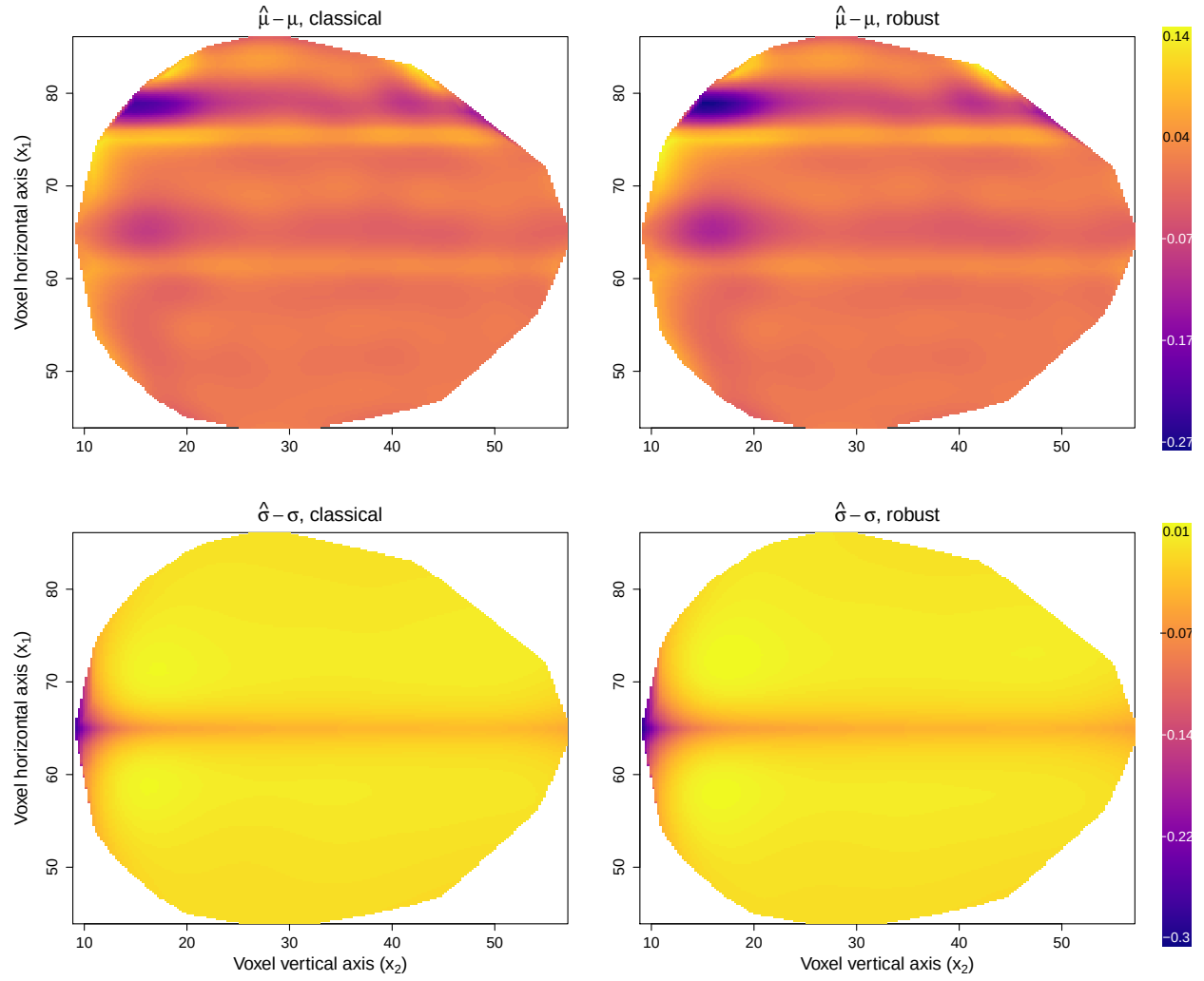


Figure S6: GAMLSS simulation, surfaces of the average bias for the canonical parameters μ (top row) and σ (bottom row) based on classical (left column) and robust (right column) estimation methods, at the assumed model.

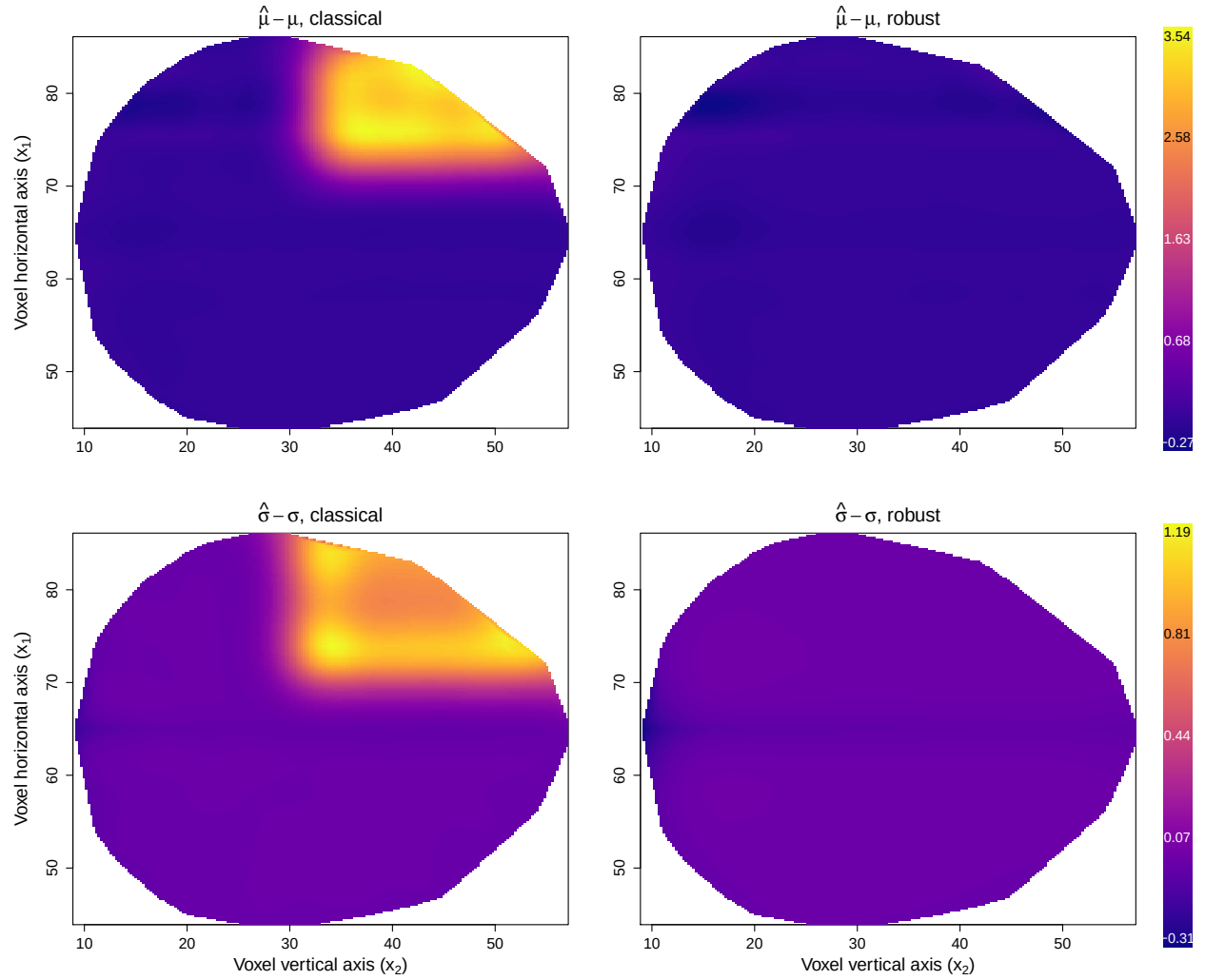


Figure S7: GAMLSS simulation, surfaces of the average bias for the canonical parameters μ (top row) and σ (bottom row) based on classical (left column) and robust (right column) estimation methods, under contamination.

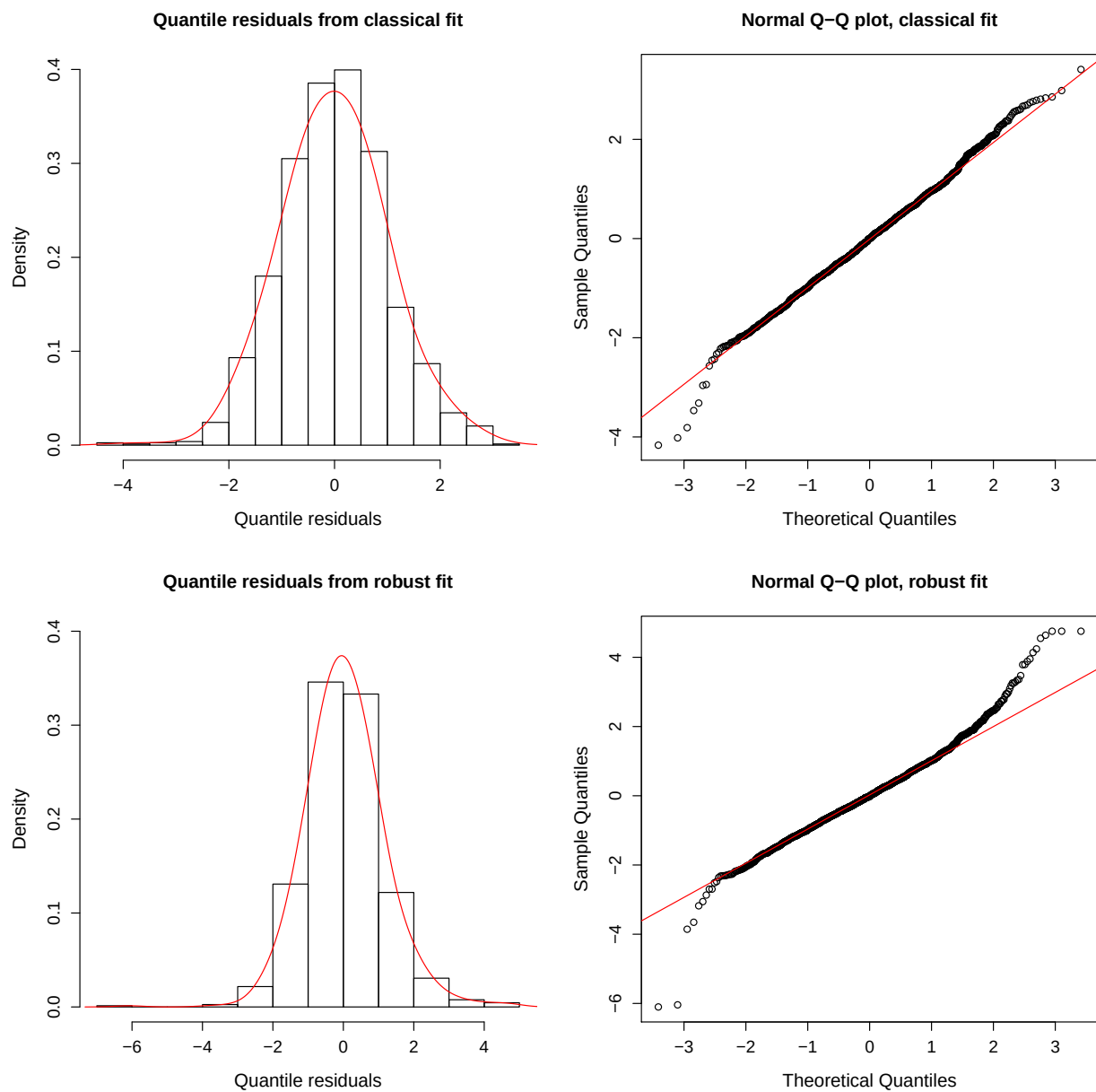


Figure S8: Brain imaging data, histogram and normal Q-Q plot of quantile residuals from the classical and robust gamma GAMLSS fits.

Web Appendix D: Supplementary Tables

Table S1 below lists the distributions implemented in the R package **GJRM**. These have been parametrized according to Rigby and Stasinopoulos (2005) and Rigby et al. (2019) and are defined in terms of μ , σ and ν . The means and variances of **DAGUM**, **FISK** (also known as log-logistic) and **SM** are indeterminate for certain values of σ and ν . If a parameter can only take positive values then the transformation/link function $\log(\cdot - \epsilon)$ is employed. If a parameter can take values in $(0, 1)$ then the inverse of the cumulative distribution function of a standardized logistic is used. $B(\cdot, \cdot)$ is the beta function, $\Gamma(\cdot)$ is the gamma function, $\alpha = \sqrt{\frac{1}{\sigma^2} + \frac{2\mu}{\sigma}}$ and $K_\lambda(t) = \frac{1}{2} \int_0^\infty x^{\lambda-1} \exp\{-0.5t(x + x^{-1})\} dx$ is the modified Bessel function of the third kind. Argument **margin** of **gamlss()** in **GJRM** allows the user to employ the desired marginal distribution and can be set to any of the values within brackets next to the names of the distributions. In many cases the parameters of the distributions determine $\mathbb{E}[Y]$ and $\mathbb{V}[Y]$ through functions thereof.

	$f(y \mu, \sigma, \nu)$	$\mathbb{E}(Y)$	$\mathbb{V}(Y)$	Support of y and Parameter ranges
beta ("BE")	$\frac{y^{\alpha_1-1}(1-y)^{\alpha_2-1}}{B(\alpha_1, \alpha_2)}$	μ	$\sigma^2 \mu(1-\mu)$	$0 < y < 1$ $0 < \mu < 1, 0 < \sigma < 1$
Dagum ("DAGUM")	$\frac{\sigma \nu}{y} \left[\frac{\left(\frac{y}{\mu}\right)^{\sigma \nu}}{\left\{\left(\frac{y}{\mu}\right)^{\sigma} + 1\right\}^{\nu+1}} \right]$	$-\frac{\mu}{\sigma} \frac{\Gamma(-\frac{1}{\sigma})\Gamma(\frac{1}{\sigma}+\nu)}{\Gamma(\nu)}$ if $\sigma > 1$	$-\left(\frac{\mu}{\sigma}\right)^2 \left[2\sigma \frac{\Gamma(-\frac{2}{\sigma})\Gamma(\frac{2}{\sigma}+\nu)}{\Gamma(\nu)} \right.$ $\left. + \left\{ \frac{\Gamma(-\frac{1}{\sigma})\Gamma(\frac{1}{\sigma}+\nu)}{\Gamma(\nu)} \right\}^2 \right]$ if $\sigma > 2$	$y > 0$ $\mu > 0, \sigma > 0, \nu > 0$
Fisk ("FISK")	$\frac{\sigma \nu^{\sigma-1}}{\mu^{\sigma} \left\{ 1 + \left(\frac{y}{\mu}\right)^{\sigma} \right\}^2} \frac{1}{y \sigma^2 - 1} \exp\left(-\frac{y}{\mu \sigma^2}\right)$	$-\frac{\mu \pi / \sigma}{\sin(\pi / \sigma)}$ if $\sigma > 1$	$\mu^2 \left\{ \frac{2\pi / \sigma}{\sin(2\pi / \sigma)} - \frac{(\pi / \sigma)^2}{\sin(\pi / \sigma)^2} \right\}$ if $\sigma > 2$	$y > 0$ $\mu > 0, \sigma > 0$
gamma ("GA")	$\frac{1}{(\mu \sigma^2)^{\frac{1}{\sigma^2}}} \frac{1}{y \sigma^2 - 1} \exp\left(-\frac{y}{\mu \sigma^2}\right) \Gamma\left(\frac{1}{\sigma^2}\right)$	μ	$\mu^2 \sigma^2$	$y > 0$ $\mu > 0, \sigma > 0$
Gumbel ("GU")	$\frac{1}{\sigma} \exp\left\{\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right)\right\}$	$\mu - 0.57722\sigma$	$\frac{\pi^2 \sigma^2}{6}$	$-\infty < y < \infty$ $-\infty < \mu < \infty, \sigma > 0$
inverse Gaussian ("IG")	$\frac{1}{\sqrt{2\pi\sigma^2 y^3}} \exp\left\{-\frac{1}{2\mu^2\sigma^2 y} (y-\mu)^2\right\}$	μ	$\mu^3 \sigma^2$	$y > 0$ $\mu > 0, \sigma > 0$
log-normal ("LN")	$\frac{1}{y\sigma\sqrt{2\pi}} \exp\left[-\frac{\{\log(y)-\mu\}^2}{2\sigma^2}\right]$	$\sqrt{\exp(\sigma^2)} \exp(\mu)$	$\exp(\sigma^2) \left\{ \exp(\sigma^2) - 1 \right\} \exp(2\mu)$	$y > 0$ $-\infty < \mu < \infty, \sigma > 0$
logistic ("LO")	$\frac{1}{\sigma} \left\{ \exp\left(-\frac{y-\mu}{\sigma}\right) \right\} \left\{ 1 + \exp\left(-\frac{y-\mu}{\sigma}\right) \right\}^{-2}$	μ	$\frac{\pi^2 \sigma^2}{3}$	$-\infty < y < \infty$ $-\infty < \mu < \infty, \sigma > 0$
normal ("N")	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$	μ	σ^2	$-\infty < y < \infty$ $-\infty < \mu < \infty, \sigma > 0$
reverse Gumbel ("RGU")	$\frac{1}{\sigma} \exp\left\{\left(-\frac{y-\mu}{\sigma}\right) - \exp\left(-\frac{y-\mu}{\sigma}\right)\right\}$	$\mu + 0.57722\sigma$	$\frac{\pi^2 \sigma^2}{6}$	$-\infty < y < \infty$ $-\infty < \mu < \infty, \sigma > 0$
Singh-Maddala ("SM")	$\frac{\sigma \nu y^{\sigma-1}}{\mu^{\sigma} \left\{ 1 + \left(\frac{y}{\mu}\right)^{\sigma} \right\}^{\nu+1}}$	$\mu \frac{\Gamma(1+\frac{1}{\sigma})\Gamma(\nu)}{\Gamma(\nu)}$ if $\sigma \nu > 1$	$\mu^2 \left\{ \Gamma\left(1+\frac{2}{\sigma}\right) \Gamma(\nu) \Gamma\left(-\frac{2}{\sigma}+\nu\right) - \Gamma\left(1+\frac{1}{\sigma}\right)^2 \Gamma\left(-\frac{1}{\sigma}+\nu\right)^2 \right\}$ if $\sigma \nu > 2$	$y > 0$ $\mu > 0, \sigma > 0, \nu > 0$
Weibull ("WEI")	$\frac{\sigma}{\mu} \left(\frac{y}{\mu}\right)^{\sigma-1} \exp\left\{-\left(\frac{y}{\mu}\right)^{\sigma}\right\}$	$\mu \Gamma\left(\frac{1}{\sigma}+1\right)$	$\mu^2 \left[\Gamma\left(\frac{2}{\sigma}+1\right) - \left\{ \Gamma\left(\frac{1}{\sigma}+1\right) \right\}^2 \right]$	$y > 0$ $\mu > 0, \sigma > 0$
Poisson ("PO")	$\frac{\exp(-\mu)\mu^y}{y!}$	μ	μ	$y = 0, 1, \dots$ $\mu > 0$
negative binomial type I ("NBI")	$\frac{\Gamma(y+1/\sigma)}{\Gamma(1/\sigma)\Gamma(y+1)} \left(\frac{\sigma\mu}{1+\sigma\mu}\right)^y \left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma}$	μ	$\mu + \sigma \mu^2$	$y = 0, 1, \dots$ $\mu > 0, \sigma > 0$
Poisson-inverse Gaussian ("PIG")	$\left(\frac{2\alpha}{\pi}\right)^{0.5} \frac{\mu^y \exp(1/\sigma) K_{y-0.5}(\alpha)}{(\alpha\sigma)^y y!}$	μ	$\mu + \sigma \mu^2$	$y = 0, 1, \dots$ $\mu > 0, \sigma > 0$
negative binomial type II ("NBII")	$\frac{\Gamma(y+\mu/\sigma)\sigma^y}{\Gamma(\mu/\sigma)\Gamma(y+1)(1+\sigma)^{y+\mu/\sigma}}$	μ	$(1+\sigma)\mu$	$y = 0, 1, \dots$ $\mu > 0, \sigma > 0$
zero-truncated Poisson ("ZTP")	$\frac{\mu^y}{y!(1-\exp(-\mu))}$	$\frac{\mu}{1-\exp(-\mu)}$	$\frac{\mu(1-\exp(-\mu)(\mu+1))}{(1-\exp(-\mu))^2}$	$y = 1, 2, \dots$ $\mu > 0$

Table S1: Definition and some properties of the distributions implemented in GJRM.

Table S2: GAMLSS simulation, summary statistics of the MSE for the classical (ML) and robust methods, on the linear predictor scale η_1 and η_2 (SD is standard deviation, IQR is inter-quartile range).

	η_1		η_2	
	Classical	Robust	Classical	Robust
At the model				
Average	0.00365	0.00393	0.00693	0.00758
SD	0.00062	0.00073	0.00215	0.00363
Median	0.00358	0.00388	0.00652	0.00713
IQR	0.00075	0.00094	0.00287	0.00288
Under 5% contamination				
Average	0.25212	0.00403	0.55575	0.00773
SD	0.00958	0.00083	0.02301	0.00232
Median	0.25198	0.00389	0.55438	0.00743
IQR	0.00832	0.00119	0.02511	0.00294

Table S3: GAMLSS simulation, summary statistics of the average bias for the classical (ML) and robust methods, on the linear predictor scale η_1 and η_2 .

	η_1		η_2	
	Classical	Robust	Classical	Robust
At the model				
Average	-0.00150	-0.00265	-0.02990	-0.03041
SD	0.02939	0.03349	0.03768	0.04431
Median	-0.00146	-0.00149	-0.02048	-0.01875
IQR	0.02432	0.02552	0.03419	0.04039
Under 5% contamination				
Average	0.21209	-0.00276	0.30637	-0.02637
SD	0.44612	0.03446	0.66538	0.04758
Median	0.00250	-0.00211	-0.03833	-0.01464
IQR	0.07547	0.02613	0.26257	0.04918

Table S4: GAM simulation, summary statistics of the average bias for the robust methods.

	AS	CGP	WYL	GAMLSS RAIC	GAMLSS RBIC	GAMLSS EFS
At the model						
Average	2.22	−0.02	−0.27	−0.05	−0.05	−0.05
SD	6.38	0.48	0.82	0.50	0.51	0.49
Median	0.06	−0.04	−0.11	−0.03	−0.05	−0.05
IQR	0.69	0.60	0.76	0.59	0.62	0.58
Under 5% contamination						
Average	1.27	−0.03	−0.31	0.01	0.01	−0.06
SD	4.91	0.54	0.95	1.11	1.14	0.88
Median	−0.02	−0.05	−0.16	−0.07	−0.08	−0.09
IQR	0.76	0.71	0.83	0.68	0.72	0.69
Under 10% contamination						
Average	1.22	0.05	−0.29	3.31 [†]	2.79 [†]	3.89 [†]
SD	4.82	0.72	1.08	34.71 [†]	28.98 [†]	39.40 [†]
Median	0.05	0.00	−0.18	−0.07	−0.10	−0.15
IQR	0.88	0.84	0.97	0.84	0.87	0.88

[†]magnitude due to six samples creating divergence

Table S5: GAM simulation, summary statistics of the timings (in seconds) for all the methods, at the model. The timings for GAMLSS RAIC and GAMLSS RBIC are for estimation with a fixed smoothing parameter obtained after a grid search, the timings of which are reported separately in the next table.

	Min	Q1	Median	Mean	Q3	Max
AS	81.436	168.667	203.438	196.055	231.169	303.375
CGP	0.296	0.609	0.750	0.973	1.009	5.785
WYL	0.807	8.438	11.647	13.108	16.463	43.728
GAMLSS RAIC	1.693	2.790	3.143	3.199	3.533	5.224
GAMLSS RBIC	1.907	2.653	2.992	3.021	3.355	5.523
GAMLSS EFS	2.574	4.996	5.455	6.297	6.070	149.442
Classical	0.045	0.055	0.062	0.065	0.071	0.325

Table S6: GAM simulation, summary statistics of the timings (in seconds) for the grid searches in GAMLSS RAIC and GAMLSS RBIC, at the model. Our searches for λ start from GAMLSS EFS and then iteratively adapts a grid of 10 equally-spaced values centered on the previous iterate. At each iteration, the grid window shrinks (by factors 0.5, 0.25, 0.125, ...) whenever the smallest value is within the interior of the grid, or the window shifts without shrinking whenever the smallest value is on the boundary. The iterations stop when any of the these conditions are satisfied: (i) the RAIC/RBIC does not change more than 10^{-5} of its previous value (relatively); (ii) no value within the current grid yields a smaller RAIC/RBIC than the previous one; (iii) the maximum number of iterations ($= 20$) is reached.

	Min	Q1	Median	Mean	Q3	Max
GAMLSS RAIC	22.404	65.340	100.902	148.535	146.981	835.102
GAMLSS RBIC	22.864	60.399	75.660	87.647	107.759	350.679

Web Appendix E: Usage of the `gamlss` function in the GJRM R Package

The proposed robust estimation for GAMLSS is readily available via the `gamlss` function in the R package GJRM (Marra and Radice, 2020). The main R steps on how to fit a robust GAMLSS are provided below.

1. Specify a list containing as many equations as the number of the distributional parameters. For example, if we consider a Gumbel distribution, then

```
eq.mu <- y ~ s(x1, x2)
eq.s2 <-    ~ s(x1, x2)
fl    <- list(eq.mu, eq.s2)
```

where `eq.mu` is the equation for μ and `eq.s2` for σ^2 , symbols `s()` refer to the unknown smooth functions as introduced in Section 2.1 in the main body of the paper, and `x1` and `x2` refer to covariates.

2. After fitting a classical GAMLSS on some data stored in a data frame `data`, the user needs to find the bounds used to approximate the integrals/sums involved in the Fisher consistency correction term $b_\rho(\delta)$ using the `rob.int` command:

```
out <- gamlss(fl, margin="GU", data=data)
bounds <- rob.int(out, rc, l.grid=1000, tol=1e-4, var.range=NULL)
lB <- bounds$rob.int$lB
uB <- bounds$rob.int$uB
```

where `rob.int` requires a fitted `gjrm` object, from a non-robust fit, a robust tuning constant (`rc`), a length of grid (`l.grid`), a numerical tolerance (`tol`), and optionally a range of values (`var.range`), to use in calculations.

3. The `rob.const` function is used to compute the median downweighting proportion (MDP) criterion corresponding to a supplied value of `rc`. We first need to fit a robust GAMLSS with tuning bounds fixed at values of `lB` and `uB` obtained in the previous step:

```
out.rpb <- gamlss(fl, margin="GU", data=data, robust=TRUE,
                 sp.method="efs", rc=rc, lB=lB, uB=uB)
rc <- rob.const(out.rpb, B=100)
```

where the `sp.method` argument to `gamlss` sets how the smoothing parameters are to be selected if not set by the user through the optional argument `sp`. The `rob.const` function requires two arguments: a fitted `gjrm` object from a robust fit and the number of bootstrap replicates (`B`). The calculations in `rob.const` encompass Steps 2–3 in Section 3.5 in the main body of the paper, and need to be repeated by some trial-and-error to find the `rc` that matches a target MDP, e.g. 0.95.

4. Once the integration bounds and the robust tuning constant have been set, the minimization of RAIC/RBIC require the construction of a grid for the smoothing parameters. For each combination of smoothing parameters, fit a robust GAMLSS and select the final one using the `rIC` function which returns both RAIC and RBIC. For example,

```
g.s <- 100
grid.search <- as.matrix(expand.grid(
  start_smo1 = seq(from=1e-7, to=0.5, length.out=g.s),
  start_smo2 = seq(from=1, to=2, length.out=g.s)
))

AICr <- NA
out.robus <- list()
for (j in 1:dim(grid.search)[1]){
  out.robus[[j]] <- gamlss(fl, margin="GU", data=data, robust=TRUE,
```

```

rc=rc, lB=lB, uB=uB, sp=c(grid.search[j,]))
AICr[j] <- rIC(out.robust[[j]])$rAIC
}
out.robust <- out.robust[[which(AICr == min(AICr))]]

```

References

- Huber, P. J. (1967). The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. Berkeley, CA: University of California Press, 221–233.
- Marra, G. and R. Radice (2020). *GJRM: Generalised Joint Regression Modelling*. R package version 0.2-3. URL: <http://CRAN.R-project.org/package=GJRM>.
- Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society, Series C* 54, 507–554.
- Rigby, R. A., M. D. Stasinopoulos, G. Z. Heller, and F. De Bastiani (2019). *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*. Boca Raton, FL: Chapman & Hall/CRC.
- Wood, S. N. and M. Fasiolo (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics* 73 (4), 1071–1081.