RECONCILING ROBUSTNESS AND FINITE-SAMPLE EFFICIENCY IN GAUSSIAN PROCESS REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust fitting reduces the influence of outliers but can sacrifice finite-sample predictive efficiency when the data follow the assumed model. In this work, we quantify how robust fitting changes the finite-sample predictive gain from Studentization, which accounts for uncertainty in an estimated variance scale rather than treating its estimate as known. Under the assumed Gaussian process (GP) model with a shared variance scale, fixed kernel parameters, and more than two training observations, Studentization yields a strictly positive finite-sample gain in expected log predictive density over plug-in Gaussian prediction, and this gain is available in closed form. We combine Studentization with density power divergence (DPD) fitting and quantify the predictive cost of robustification. Our analysis relates this cost to model complexity and the influence of the training responses on test predictions. It gives a sufficient condition for retaining a prescribed fraction of the exact Studentization gain over ordinary Gaussian prediction. Controlled synthetic experiments support the analysis and show predominantly positive gains from Studentization in prediction. Beyond the theoretical assumptions, applying Studentization to existing robust GP methods mostly improves joint predictive log likelihood, with only small decreases otherwise. Finally, we derive a distribution-free bound on the possible decrease in log predictive density when Studentization is applied to an already fitted model. The code is available at https://anonymous.4open.science/r/rsgpr-0E17/.

1 INTRODUCTION

Predictive models should balance finite-sample efficiency under the assumed model with robustness against outliers. Robust fitting provides protection against outliers, but can incur an efficiency cost when the data follow the assumed model (Huber, 1964; Basu et al., 1998). We assess this cost relative to the finite-sample predictive gain from Studentization, which accounts for uncertainty in an estimated variance scale rather than treating the scale estimate as known and, in our setting, yields a strictly positive gain in expected log predictive density over plug-in Gaussian prediction. This tradeoff is especially relevant in probabilistic regression, where robustness should protect the fitted model without unnecessarily degrading the predictive distribution on clean data.

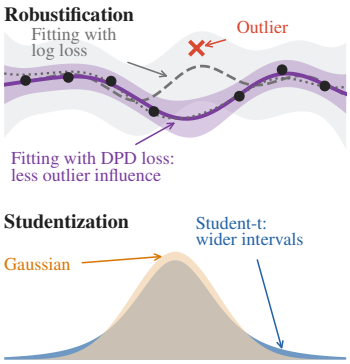


Figure 1: Robust fitting reduces the outlier's influence on the fitted curve (top). Studentized predictions give wider intervals (bottom).

We study a setting in which this gain is available in closed form. In the Gaussian process (GP) model considered here, the prior covariance and observation noise variance share a precision τ (variance scale τ^{-1}) with prior $\pi(\tau) \propto \tau^{-1}$, and integrating out τ gives Studentized predictions with exactly n degrees of freedom, where n is the training sample size. Under the assumed model with fixed kernel parameters and $n > 2$, their gain over the ordinary Gaussian predictor, measured by expected log predictive density, is the exact quantity $\Delta_{n,p} = D_{\mathrm{KL}}(\mathrm{St}_n^{(p)} \,\|\, \varphi_p) > 0$ (Lemma 1).

The exact gain separates into two parts. Writing Δ_n for the marginal gain at a single test input, Lemma 1 gives $\Delta_{n,p} = p\Delta_n + \mathcal{I}_{n,p}$ for p jointly predicted test targets. The second term is the excess

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

of the joint gain over the sum of marginal gains. It arises because one scale is shared across test inputs, so integrating it out adds dependence beyond that encoded by the GP covariance; a plug-in Gaussian predictor, treating the estimated precision as known, omits this additional dependence. This distinction matters whenever predictions are evaluated jointly rather than one test input at a time, because predictive uncertainty then includes both marginal behavior and dependence across targets. Joint prediction is thus the setting in which Studentization does more than reshape marginals, and we take the joint predictive log likelihood (joint PLL), evaluated as the log predictive density, as the primary metric throughout.

Robust GP methods modify the observation model or fitting criterion to reduce outlier influence (Jylänki et al., 2011; Altamirano et al., 2024; Ament et al., 2024). More broadly, robust learning methods typically modify how the model is fitted, whereas predictive uncertainty is determined by the distribution used at prediction time. Although robust fitting and Studentization act at different stages, their effects are coupled through the fitted mean, predictive scale, and precision estimate (Figure 1). Thus, the gain available under the ordinary fit need not survive robustification. This paper asks a single fundamental question: when does Studentization help, and when does it not? We organize the theoretical and empirical results along three axes, namely the prediction dimension p, the training sample size n, and the contamination pattern.

We fit the model using a density power divergence (DPD) loss within generalized variational inference (GVI) (Basu et al., 1998; Bissiri et al., 2016; Knoblauch et al., 2022), where the robustification parameter β controls the degree of robustification and DPD downweights outliers for $\beta > 0$. At prediction time, a Kullback–Leibler (KL) projection fixes the mean precision at its robust estimate while retaining uncertainty in this parameter, and integrating out the projected precision gives Studentized predictions with n degrees of freedom. At $\beta = 0$ the construction recovers ordinary GP fitting and ordinary Studentization, providing the exact baseline of Lemma 1.

Our contributions are as follows.

- Lemma 1 gives the Studentization gain $\Delta_{n,p}$ at $\beta = 0$ in closed form, together with the separation $\Delta_{n,p} = p\Delta_n + \mathcal{I}_{n,p}$ that identifies the dependence term available only for $p > 1$.
- For the fitted solutions constructed in Appendix E.2, Theorem 1 expands the gain of the DPD Studentized predictor over the ordinary Gaussian predictor as $G_p(\beta) = \Delta_{n,p} - c_{1,p}\beta + r_p(\beta)$ with $|r_p(\beta)| \leq \overline{C}_{n,p}\beta^2$ on $[0, b_0]$. The first-order coefficient $c_{1,p}$ increases with model complexity and with the influence of the training responses on test predictions, and the expansion yields a sufficient condition for retaining any prescribed fraction of $\Delta_{n,p}$.
- For any fitted mean and positive definite predictive scale, Proposition 1 bounds the possible PLL decrease uniformly, without assumptions on the data distribution or the fitting procedure.
- Controlled synthetic experiments validate the expansion and assess gain retention; injecting outliers separates the contribution of DPD fitting from that of Studentization; and applying Studentization after fitting three existing robust GP models on synthetic and real data tests the rule outside the assumptions of the DPD analysis.

2 Related work

Studentization and predictive Student-t distributions. In Gaussian models with an unknown variance scale, Studentization can be expressed by integrating out that scale, thereby generating Student-t predictions that account for the uncertainty omitted by a plug-in Gaussian predictor. Predictive density estimation provides a framework for comparing such predictions through measures such as information loss (Aitchison, 1975) and KL risk (Liang & Barron, 2004). In spatial prediction, Handcock & Stein (1993) account for the uncertainty in covariance parameters within the Bayesian kriging framework. Student-t processes offer a related construction at the level of the function prior, obtained by integrating an inverse Wishart process prior over the covariance kernel (Shah et al., 2014). By using a scale-invariant precision prior $\pi(\tau) \propto \tau^{-1}$, predictions from our shared-scale Gaussian model exactly follow a Student-t distribution with n degrees of freedom. After standardizing the GP predictions by the predictive scale, integrating out the shared precision induces dependence among prediction errors that a plug-in Gaussian predictor omits; this is the term $\mathcal{I}_{n,p}$ isolated in Lemma 1. This exact gain relative to the ordinary Gaussian predictor provides an analytically tractable baseline for evaluating the predictive cost of robustification.

Robust GP regression. Student-t likelihoods use the same distributional family to mitigate the impact of outliers during fitting, but this plays a different role than integrating out the scale at prediction time. Jylänki et al. (2011) study approximate inference for GPs with Student-t observations, while Tang et al. (2017) combine such observations with a Student-t process prior. Other robust GP methods involve modifying the fitting criterion or noise structure. For example, RCGP uses generalized Bayesian inference while maintaining conjugacy (Altamirano et al., 2024), and Relevance Pursuit introduces additional noise to individual observations (Ament et al., 2024). Lu et al. (2023) model inliers as Gaussian distributions and outliers as uniform distributions via a mixture likelihood, fitting the model using variational inference. These last three methods form the final stage of our evaluation. They provide explicit robust fitting procedures with learned kernel parameters, allowing Studentization to be assessed after fitting while keeping the fitted mean and predictive scale fixed. This enables verification beyond the assumptions of our DPD analysis.

3 Preliminaries

3.1 Studentization and finite-sample prediction

Let $\tau_0 > 0$ be the unknown precision and $\widehat{\tau}$ its estimate from training data. Suppose the scalar prediction target y satisfies $\mathrm{z} = \sqrt{\tau_0/v}(\mathrm{y} - \mu) \sim \mathcal{N}(0,1)$, independent of the scale estimate $\mathrm{w} = \nu\tau_0/\widehat{\tau} \sim \chi^2(\nu)$, for $\mu \in \mathbb{R}$, $v > 0$, and $\nu > 0$. Studentization standardizes the error; across repeated samples, it follows a Student-t distribution:

$$\mathrm{t} = \frac{\mathrm{y} - \mu}{\sqrt{v/\widehat{\tau}}} = \sqrt{\frac{\nu}{\mathrm{w}}}\,\mathrm{z} \sim \mathrm{St}_\nu(\mathrm{t}; 0, 1). \tag{1}$$

Consequently, the ordinary Studentized predictor $q_{\mathrm{St}}(\mathrm{y} \mid \widehat{\tau}) = \mathrm{St}_\nu(\mathrm{y}; \mu, v/\widehat{\tau})$ strictly reduces the expected negative log predictive density $\mathcal{L}_{\log}(q) = \mathbb{E}[-\log q(\mathrm{y} \mid \widehat{\tau})]$ relative to the ordinary Gaussian predictor $q_{\mathcal{N}}(\mathrm{y} \mid \widehat{\tau}) = \mathcal{N}(\mathrm{y}; \mu, v/\widehat{\tau})$ for $\nu > 2$:

$$\mathcal{L}_{\log}(q_{\mathcal{N}}) - \mathcal{L}_{\log}(q_{\mathrm{St}}) = \mathbb{E}\left[\log \frac{\sqrt{\widehat{\tau}/v}\,\mathrm{St}_\nu(\mathrm{t}; 0, 1)}{\sqrt{\widehat{\tau}/v}\,\mathcal{N}(\mathrm{t}; 0, 1)}\right] = D_{\mathrm{KL}}(\mathrm{St}_\nu(\cdot; 0, 1) \,\|\, \mathcal{N}(\cdot; 0, 1)) > 0. \tag{2}$$

3.2 GP regression and Studentization

Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be the training data with inputs $\boldsymbol{X}$ and responses $\mathbf{y}$. For a latent function f and positive definite kernel k, we consider a GP model where the latent covariance k/τ and observation noise variance σ^2/τ share the precision parameter $\tau > 0$:

$$\begin{aligned} \pi(\tau) &\propto \tau^{-1}, \\ \mathrm{f} \mid \tau &\sim \mathcal{GP}(\mathrm{f}; 0, k/\tau), \\ \mathrm{y}_i \mid \mathrm{f}_i, \tau &\sim \mathcal{N}(\mathrm{y}_i; \mathrm{f}_i, \sigma^2/\tau). \end{aligned} \tag{3}$$

Define $\boldsymbol{K}_y = k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}_n$ and $Q_{\mathbf{y}} = \mathbf{y}^\top \boldsymbol{K}_y^{-1} \mathbf{y}$. For a test input $\boldsymbol{x}_*$, let $v_* = \widetilde{k}_{**} - k(\boldsymbol{x}_*, \boldsymbol{X})\boldsymbol{K}_y^{-1} k(\boldsymbol{X}, \boldsymbol{x}_*)$, where $\widetilde{k}_{**}$ includes σ^2 for noisy response prediction. Conditioned on τ, the prediction is Gaussian with mean $\mu_*^{(0)}(\mathbf{y})$ and variance v_*/τ (Rasmussen & Williams, 2006, Chapter 2). Substituting the maximum likelihood estimate $\widehat{\tau}^{(0)}(\mathbf{y}) = n/Q_{\mathbf{y}}$ gives the ordinary Gaussian predictor. Integrating the conditional prediction against the Gamma posterior $p(\tau \mid \mathbf{y}) = \mathrm{Gam}(\tau; n/2, Q_{\mathbf{y}}/2)$ is exactly equivalent to Studentization with $\nu = n$:

$$q_0(\mathrm{y}_* \mid \mathbf{y}) = \int_0^\infty p(\mathrm{y}_* \mid \tau, \mathbf{y}) p(\tau \mid \mathbf{y})\, d\tau = \mathrm{St}_n\left(\mathrm{y}_*; \mu_*^{(0)}(\mathbf{y}), Q_{\mathbf{y}} v_*/n\right). \tag{4}$$

3.3 GVI and DPD

GVI allows the data fitting loss and the regularizer toward the prior to be chosen separately (Bissiri et al., 2016; Knoblauch et al., 2022). For a loss $\ell(\boldsymbol{\theta}, y)$, prior $\pi(\boldsymbol{\theta})$, and variational family $\mathcal{Q}$, choosing a KL regularizer gives

$$q^* = \arg\min_{q \in \mathcal{Q}} \left\{ \sum_{i=1}^n \mathbb{E}_{q(\boldsymbol{\theta})}[\ell(\boldsymbol{\theta}, y_i)] + D_{\mathrm{KL}}(q(\boldsymbol{\theta}) \,\|\, \pi(\boldsymbol{\theta})) \right\}.$$

The negative log likelihood $\ell(\boldsymbol{\theta}, y) = -\log p(y; \boldsymbol{\theta})$ recovers ordinary variational inference.

To reduce outlier influence, we instead use the DPD loss for $\beta > 0$:

$$S_\beta(p, y) = \frac{1}{1+\beta} \int p(z; \boldsymbol{\theta})^{1+\beta} \, dz - \frac{1}{\beta} p(y; \boldsymbol{\theta})^\beta + \frac{1}{\beta} - 1, \tag{5}$$

which recovers the log loss $S_0 = -\log p(y)$ as $\beta \to 0$. The density power term bounds the loss contribution of observations with very low model density (outliers) (Basu et al., 1998). Section 4.1 applies this loss within GVI to obtain the GP fitting objective.

4 DPD FITTING AND STUDENTIZED PREDICTION

We formulate a two step procedure to combine outlier robustness with exact Studentization: (1) robust parameter estimation using DPD, and (2) a KL projection that fixes the mean precision while retaining uncertainty in this parameter.

4.1 DPD OBJECTIVE

We approximate the conditional posterior of the latent function $\mathbf{f}$ given τ by a Gaussian $q(\mathbf{f} \mid \tau) = \mathcal{N}(\mathbf{f}; \boldsymbol{m}, \boldsymbol{S}/\tau)$. Let $v_i = [\boldsymbol{S}]_{ii}$, $e_i = y_i - m_i$, and $d_i = \sigma^2 + \beta v_i$. Averaging the DPD loss analytically under the Gaussian variational distribution gives the observation weights:

$$w_i = \sqrt{\sigma^2/d_i} \exp\left(-\beta \tau e_i^2/(2d_i)\right). \tag{6}$$

The weight w_i decays exponentially for large standardized residuals, so the DPD loss automatically downweights outliers when $\beta > 0$.

Adding a KL penalty relative to the prior (Bissiri et al., 2016) gives our fitting objective:

$$J_\beta(\boldsymbol{m}, \boldsymbol{S}, \tau) = \zeta_\beta(\tau) \sum_{i=1}^{n} \left(\xi_\beta - \frac{w_i}{\beta} \right) + \frac{n}{\beta} - n + D_{\mathrm{KL}}(q(\mathbf{f} \mid \tau) \,\|\, p(\mathbf{f} \mid \tau)), \tag{7}$$

where $\zeta_\beta(\tau) = (\tau/(2\pi\sigma^2))^{\beta/2}$ and $\xi_\beta = (1+\beta)^{-3/2}$. We minimize J_β to obtain the selected solution $(\boldsymbol{m}^{(\beta)}, \boldsymbol{S}^{(\beta)}, \widehat{\tau}^{(\beta)})$, which recovers ordinary GP fitting at $\beta = 0$.

4.2 PRECISION PROJECTION AND STUDENTIZED PREDICTION

Stationarity of J_β with respect to log precision yields $\widehat{\tau}^{(\beta)} = \Omega_\beta / Q_\beta$, where Ω_β and Q_β aggregate the effective weights and residuals (equation 21). For prediction, we constrain the projected precision distribution to have this robust mean $\widehat{\tau}^{(\beta)}$ while minimizing the KL divergence from the precision posterior under the Gaussian model, $p(\tau \mid \mathbf{y}) = \mathrm{Gam}(\tau; n/2, Q_{\mathbf{y}}/2)$. The unique solution is:

$$q_\beta^{(\tau)}(\tau \mid \mathbf{y}) = \mathrm{Gam}\left(\tau; \frac{n}{2}, \frac{n}{2\widehat{\tau}^{(\beta)}}\right). \tag{8}$$

By retaining the shape parameter $n/2$ of the Gamma distribution, this projection preserves the relative variance $2/n$. The predictive distribution therefore has the same tail decay as a Student-t distribution with n degrees of freedom, as in ordinary Studentization.

Let $\boldsymbol{a}_* = \boldsymbol{K}^{-1}\boldsymbol{k}_*$ and $r_* = \widetilde{k}_{**} - \boldsymbol{k}_*^\top \boldsymbol{K}^{-1} \boldsymbol{k}_*$. Set $\mu_*^{(\beta)} = \boldsymbol{a}_*^\top \boldsymbol{m}^{(\beta)}$ and $v_*^{(\beta)} = r_* + \boldsymbol{a}_*^\top \boldsymbol{S}^{(\beta)} \boldsymbol{a}_*$. Integrating the Gaussian prediction against $q_\beta^{(\tau)}$ then yields the DPD Studentized predictor: $q_\beta(\mathrm{y}_* \mid \mathbf{y}) = \mathrm{St}_n\left(\mathrm{y}_*; \mu_*^{(\beta)}, v_*^{(\beta)}/\widehat{\tau}^{(\beta)}\right)$.

5 FINITE-SAMPLE PREDICTIVE PERFORMANCE

Robust fitting inevitably trades some predictive efficiency on uncontaminated data for outlier protection. We quantify this cost via a local expansion around $\beta = 0$ to obtain a sufficient condition for retaining the Studentization gain.

We consider how the fitted parameters change with β while holding the training data and kernel parameters fixed. Appendix E.2 constructs a family of fitted solutions that varies smoothly with β near zero and recovers the ordinary GP fit at $\beta = 0$. We call this family of fitted solutions the *solution path*.

5.1 PREDICTION SETUP

Assume $n > 2$ and that the training responses $\mathbf{y}$ and p jointly predicted test targets $\mathbf{y}_*$ follow the assumed GP model without outliers (precision τ_0). The true conditional covariance is $\boldsymbol{V}_*/\tau_0$, where $\boldsymbol{V}_* = \widetilde{\boldsymbol{K}}_{**} - \boldsymbol{K}_{X*}^\top \boldsymbol{K}_y^{-1} \boldsymbol{K}_{X*}$. Defining $\boldsymbol{A}_* = \boldsymbol{K}_{X*}^\top \boldsymbol{K}^{-1}$ and $\boldsymbol{R}_* = \widetilde{\boldsymbol{K}}_{**} - \boldsymbol{K}_{X*}^\top \boldsymbol{K}^{-1} \boldsymbol{K}_{X*}$, we write the predictive parameters as:

$$\text{Ordinary GP } (\beta = 0): \quad \boldsymbol{\mu}_*^{(0)} = \boldsymbol{K}_{X*}^\top \boldsymbol{K}_y^{-1} \mathbf{y}, \quad \boldsymbol{\Sigma}_*^{(0)} = Q_{\mathbf{y}} \boldsymbol{V}_*/n$$
$$\text{DPD GP } (\beta > 0): \quad \boldsymbol{\mu}_*^{(\beta)} = \boldsymbol{A}_* \boldsymbol{m}^{(\beta)}, \quad \boldsymbol{\Sigma}_*^{(\beta)} = (\boldsymbol{R}_* + \boldsymbol{A}_* \boldsymbol{S}^{(\beta)} \boldsymbol{A}_*^\top)/\widehat{\tau}^{(\beta)}.$$

We define the predictive risk $\mathcal{R}_p(q)$ as the expected KL divergence from the true conditional Gaussian density to the predictive density q. A reduction in risk corresponds exactly to an increase in expected log predictive density (Gneiting & Raftery, 2007, Section 4).

5.2 STUDENTIZATION GAIN AND DPD PREDICTIVE COST

Lemma 1 (Exact Studentization gain at $\beta = 0$)**.** *The gain of the ordinary Studentized predictor $q_0^{(p)}$ over the Gaussian predictor $q_{\mathrm{G}}^{(p)}$ is exactly*

$$\Delta_{n,p} = D_{\mathrm{KL}}(\mathrm{St}_n^{(p)} \,\|\, \varphi_p) > 0.$$

After standardization, this splits into individual marginal gains and their dependence:

$$\Delta_{n,p} = p\Delta_n + \mathcal{I}_{n,p}, \qquad \mathcal{I}_{n,1} = 0.$$

To quantify the cost of DPD fitting, define the effective degrees of freedom $\nu_{\mathrm{sm}} = \mathrm{Tr}(\boldsymbol{K}\boldsymbol{K}_y^{-1})$ and the contribution from the test inputs $h(\boldsymbol{X}_*) = \sigma^2 \,\mathrm{Tr}\big(\boldsymbol{V}_*^{-1} \boldsymbol{K}_{X*}^\top \boldsymbol{K}_y^{-2} \boldsymbol{K}_{X*}\big)$. The coefficient of the linear predictive cost is:

$$c_{1,p} = \frac{p\nu_{\mathrm{sm}}/n + h(\boldsymbol{X}_*)}{2(n+p)}. \tag{9}$$

The coefficient $c_{1,p}$ captures the robustification tradeoff. It increases with model complexity (ν_{sm}/n) and with the influence of the training responses on prediction, normalized by the conditional covariance ($h(\boldsymbol{X}_*)$).

Theorem 1 (DPD fitting and predictive gain)**.** *There exists a universal constant $c_0 > 0$ defining an interval $[0, b_0]$ independent of the responses, where $b_0 = c_0/(n + 1 + \max_i K_{ii}/\sigma^2)$. On the solution path constructed in Appendix E.2, the DPD Studentized predictor has the following gain over the ordinary Gaussian predictor, measured by expected log predictive density:*

$$G_p(\beta) = \mathcal{R}_p(q_{\mathrm{G}}^{(p)}) - \mathcal{R}_p(q_\beta^{(p)}) = \Delta_{n,p} - c_{1,p}\beta + r_p(\beta), \tag{10}$$

where $|r_p(\beta)| \le \overline{C}_{n,p}\beta^2$ for $0 \le \beta \le b_0$. Consequently, for any prescribed fraction $\rho \in (0,1)$, ensuring $c_{1,p}\beta + \overline{C}_{n,p}\beta^2 \le \rho\Delta_{n,p}$ guarantees that $G_p(\beta) \ge (1-\rho)\Delta_{n,p} > 0$.

Thus, locally, robustification incurs a linear predictive cost relative to the exact Studentization gain, up to a bounded quadratic remainder. The theorem gives a sufficient condition for retaining a prescribed fraction of $\Delta_{n,p}$; its constants are not evaluated numerically.

For predictions from an already fitted model, the following bound holds without assumptions on the data distribution or fitting procedure.

Proposition 1 (Bound on the PLL decrease). *Let* $q_{\mathrm{G}} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $q_{\mathrm{St}} = \mathrm{St}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *on* $\mathbb{R}^p$, *with* $n > 2$ *and* $\boldsymbol{\Sigma} \succ 0$. *Then*

$$\frac{1}{p} \log \frac{q_{\mathrm{St}}(\mathbf{y})}{q_{\mathrm{G}}(\mathbf{y})} \geq -\varepsilon_{n,p} \qquad \text{for every } \mathbf{y} \in \mathbb{R}^p,$$
$$\varepsilon_{n,p} = \frac{1}{p} \left[\log \frac{\Gamma(n/2)}{\Gamma((n+p)/2)} + \frac{p}{2} \log \frac{n}{2} - \frac{p}{2} + \frac{n+p}{2} \log \left(1 + \frac{p}{n}\right) \right]. \tag{11}$$

The change in mean pointwise PLL is bounded below by $-\varepsilon_{n,1}$.

Appendix G.7 gives the proof and bounds for the experimental settings. This result bounds the possible decrease in PLL; it does not guarantee improvement or nominal coverage.

6 Experiments

We design two controlled synthetic experiments to empirically connect our local risk expansion with the practical tradeoff between robustness and predictive performance. First, we validate the theoretical predictive cost on uncontaminated data. Second, we add outliers to the training data to compare the gain retained on clean data against actual robust performance while isolating the individual impacts of DPD fitting and Studentization. We then evaluate Studentization after fitting three existing robust GP models on synthetic and real data.

6.1 Setup

Our primary evaluation metric is the joint PLL of responses at p test inputs. Gains are reported as differences in PLL (in nats), with the Gaussian baseline specified for each comparison. For the two controlled experiments, synthetic data were generated from a zero mean GP with a radial basis function (RBF) kernel. These experiments covered small to moderately large training sample sizes ($n \in \{4, \ldots, 64\}$) and prediction dimensions ($p \in \{1, \ldots, 128\}$), with 500 independent datasets per condition. Numerical protocols, optimization diagnostics, and full results are given in Appendix F.

6.2 Prediction on uncontaminated data

We assess how much of the exact Studentization gain $\Delta_{n,p}$ is retained when the data follow the assumed GP model. For positive β, we estimated the predictive cost of DPD fitting by Monte Carlo, using a control variate to reduce variance (Appendix F.2.1). Subtracting this estimate from the exact Studentization gain $\Delta_{n,p}$ gave the estimated risk reduction $G_p(\beta)$.

Figure 2 compares the estimated gains with the first-order approximation from our theory ($\Delta_{n,p} - c_{1,p}\beta$). As predicted by Theorem 1, the linear penalty accurately tracks the predictive cost near $\beta = 0$. At a small β value ($\beta = 0.01$), the estimated gains remained strictly positive across all 56 (n, p) conditions. As β increases, higher order terms emerge and eventually drive the gain negative when predicting few test values (small p) with larger training sets ($n = 64$).

To assess gain retention at finite β, we estimated a correction based on second derivatives (Appendix F.2.3). We then identified conditions predicted by that estimate to retain at least 50% of the exact gain. Of the 504 conditions tested with $\beta > 0$, 446 fell within these empirical cutoffs, and independent Monte Carlo estimates exceeded 50% retention in all 446 conditions.

6.3 Prediction after adding outliers

We evaluate prediction performance after injecting outliers into the training responses while holding the kernel, input locations, and model for the test responses fixed. We corrupted a target fraction $r = 0.1$ of the training data with large uniform shifts of 3 to 9 times the standard deviation (SD) of the uncontaminated training responses (Altamirano et al., 2024).

To compare the gain retained on clean data with robustness after adding outliers, we evaluated identical training datasets before and after adding outliers at the same β values ($\beta \in \{0.01, 0.05, 0.1\}$). We decomposed the total joint PLL gain over ordinary Gaussian prediction into two sequential increments: (1) DPD Gaussian minus ordinary Gaussian, and (2) DPD Studentized minus DPD Gaussian.

Figure 3 summarizes this decomposition alongside the fraction of the gain retained on uncontaminated data. At a small β value ($\beta = 0.01$), the DPD Studentized predictors retained nearly all the exact

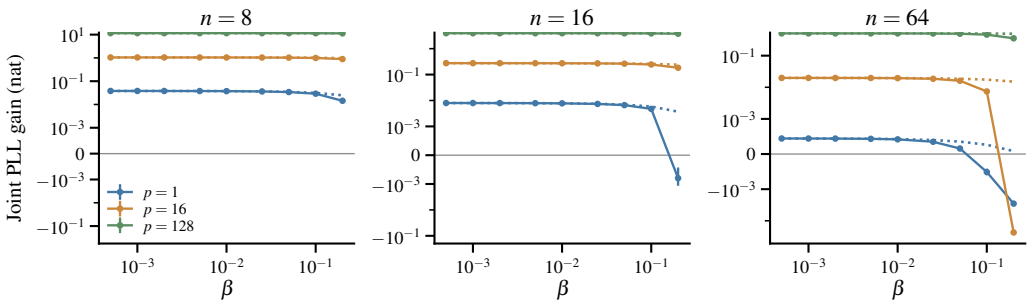


Figure 2: Gain in joint PLL over ordinary Gaussian prediction on uncontaminated data. The first-order approximation (dotted) tracks the cost near $\beta = 0$; larger β produces departures that can reverse the net gain, particularly at small p. Solid curves indicate Monte Carlo estimates with error bars showing one standard error (SE).

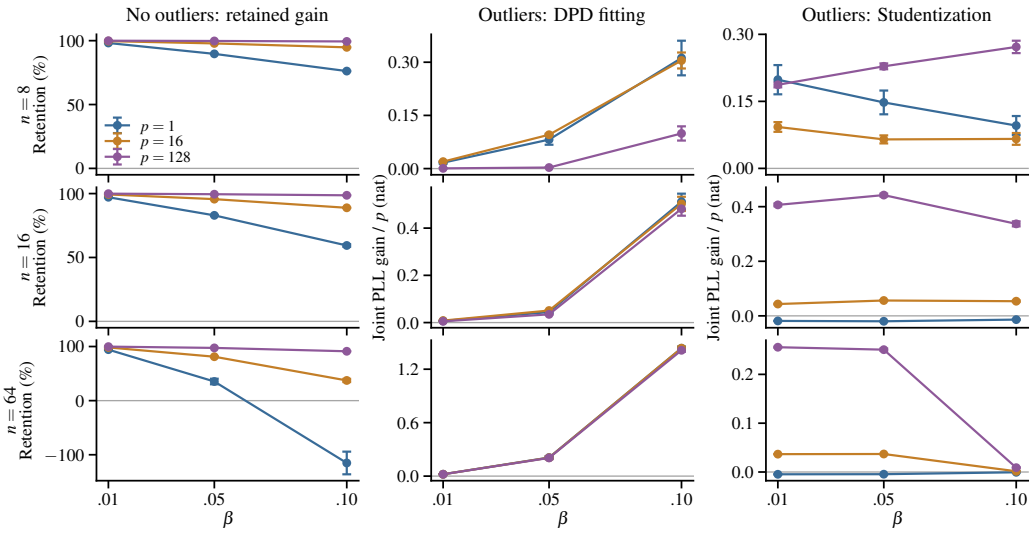


Figure 3: Tradeoff between the gain retained on clean data and predictive performance after adding outliers to the training data. Left: Retained fraction of $\Delta_{n,p}$ on clean data. Middle: Joint PLL increment from applying DPD fitting (Gaussian shape) on contaminated data. Right: Additional joint PLL increment from Studentization after fitting on contaminated data. The total gain per response is the sum of the middle and right columns. Error bars denote one Monte Carlo SE.

Studentization gain on clean data ($> 94\%$) while improving the joint PLL on contaminated data across all (n, p) conditions tested. However, the interaction between robust fitting and Studentization is more complex at larger β values and for some choices of test inputs. For instance, at $n = 64, p = 16, \beta = 0.05$, DPD fitting increased the test PLL by ~ 3.3 nats, and Studentization then added ~ 0.6 nats. Conversely, for univariate predictions ($p = 1$), Studentization frequently reduced the mean test PLL after robust fitting, even when the fraction of the exact Studentization gain retained on clean data was high.

Supplementary analyses (Appendices F.3.3–F.3.5) show that the additional gain from Studentization is small per response at the largest training sample size, but does not decrease monotonically with n, and that Studentization can reduce PLL under severe contamination (e.g., $r = 0.2$ or focused contamination of both inputs and responses).

6.4 Studentization of existing robust GPs

We evaluate whether Studentization improves predictions from existing robust GP models after fitting. We consider robust and conjugate GP regression (RCGP; Altamirano et al. 2024), Relevance Pursuit (Ament et al., 2024), and the Gaussian/uniform mixture GP of Lu et al. (2023). For each fitted model, we replace its Gaussian predictive distribution with a Student-t distribution. The degrees of freedom equal the training sample size n, and the fitted mean and predictive scale are retained without further

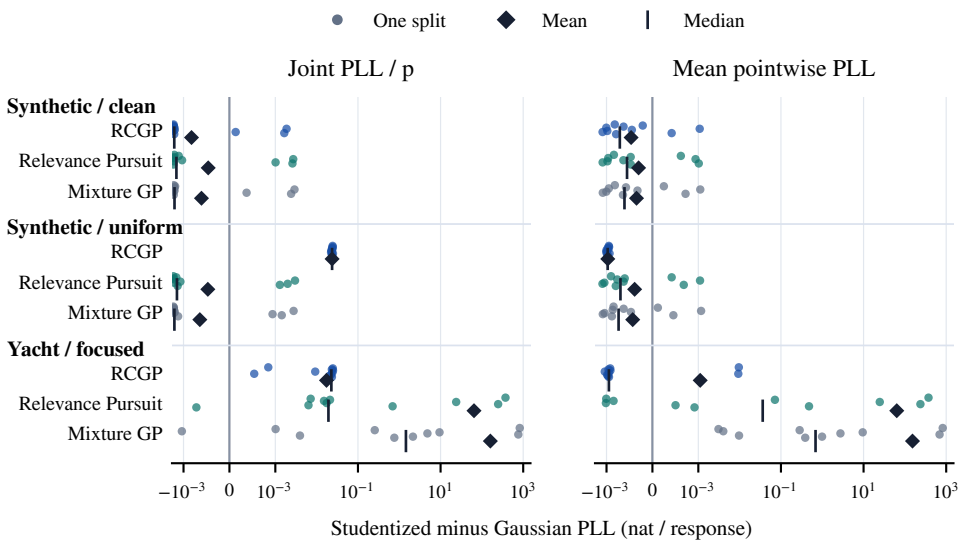


Figure 4: Paired changes in joint PLL per response (left) and mean pointwise PLL (right) in three diagnostic conditions. Each colored point is one of ten splits; black diamonds and vertical marks show the mean and median. Positive values favor Studentization; the shared axis is linear within 0.001 nat per response of zero and logarithmic outside.

optimization. Here, $\boldsymbol{\Sigma}$ is the Student-t scale matrix, so its covariance is $n\boldsymbol{\Sigma}/(n-2)$. This experiment tests a prediction rule motivated by the separation of robust fitting and Studentization, beyond the assumptions of our DPD analysis.

We used one synthetic dataset and three real datasets, Boston (Harrison & Rubinfeld, 1978), Energy (Tsanas & Xifara, 2012), and Yacht (Gerritsma et al., 1981), with clean training data and three forms of training contamination. The contamination mechanisms follow the released RCGP code, with a target contamination fraction of 10% and no injected test outliers. Each of the 16 dataset/contamination conditions was evaluated on ten 90/10 train/test splits. We compared Studentized and Gaussian predictions from the same fit using both the mean pointwise PLL and the joint PLL divided by the number of test observations. Appendix G gives the model settings, full results, and numerical diagnostics.

Studentization improved the joint PLL in most conditions for all three models (Table 1). Pointwise improvements were also frequent for Relevance Pursuit and the mixture GP, whereas RCGP more often showed small pointwise decreases. Across both metrics and all three models, the largest decrease in a condition mean was less than 0.0011 nat per response. The median joint gain across all 16 condition means was positive for each model. Figure 4 shows individual splits in three diagnostic conditions; the large gains on Yacht with focused contamination illustrate how means can differ from typical split outcomes. For the training sample sizes and prediction dimensions used here, Proposition 1 gives an upper bound of less than 0.002 nat per response on the decrease for both mean pointwise PLL and joint PLL per response. These results support Studentization as a simple way to improve joint prediction from fitted robust GPs, with small losses in the conditions where it did not improve PLL.

6.5 Discussion

We return to the question of when Studentization helps, organizing the evidence along the three axes introduced in Section 1: prediction dimension, training sample size, and contamination pattern.

Prediction dimension p. Lemma 1 gives $\mathcal{I}_{n,1} = 0$, so at a single test input Studentization only reshapes the marginal, with exact gain $\Delta_n > 0$ at $\beta = 0$ under the assumed model. In the contaminated-data experiments, for univariate predictions ($p = 1$), Studentization frequently reduced the mean test PLL relative to Gaussian predictions from the same robust fit, even when the fraction of the exact Studentization gain retained on clean data was high. For $p > 1$ the dependence term is available, and joint PLL evaluates both that dependence and the marginal densities. At $n = 64$,

Table 1: Paired changes in PLL, in 10^{-3} nats per response. Median and worst change are computed over all 16 condition means, each based on ten splits. Positive counts give the numbers of condition means and differences within individual splits above zero.

		Positive difference		All 16 condition means	
Metric	Model	Conditions	Splits	Median	Worst
Joint PLL / p	RCGP	15/16	134/160	19.344	–0.826
	Relevance Pursuit	14/16	108/160	10.053	–0.470
	Mixture GP	13/16	105/160	14.822	–0.668
Mean pointwise PLL	RCGP	6/16	41/160	–0.394	–1.035
	Relevance Pursuit	13/16	94/160	3.199	–0.387
	Mixture GP	13/16	120/160	9.762	–0.429

$p = 16$, $\beta = 0.05$, for instance, DPD fitting increased the test PLL by $\sim$3.3 nats and Studentization then added $\sim$0.6 nats.

Training sample size n. The supplementary analyses (Appendices F.3.3–F.3.5) show that the additional gain from Studentization is small per response at the largest training sample size, but it does not decrease monotonically with n. For fixed p, the exact gain $\Delta_{n,p}$ under the assumed model tends to zero as n grows (Appendix C.2). The interval $[0, b_0]$ on which Theorem 1 controls the remainder has $b_0 = c_0/(n + 1 + \max_i K_{ii}/\sigma^2)$, which shrinks with n when $\max_i K_{ii}/\sigma^2$ is fixed. A sufficient condition for retaining at least a fraction $1 - \rho$ of the exact gain within this interval is $c_{1,p}\beta + \overline{C}_{n,p}\beta^2 \leq \rho\Delta_{n,p}$, where $\rho \in (0, 1)$. The same expansion also indicates that an overly flexible model raises $c_{1,p}$ through ν_{sm}/n, so the penalty from DPD fitting can exceed the exact gain as β increases.

Contamination pattern. Retaining the Studentization gain on clean data does not guarantee that Studentization will improve performance on contaminated data. The additional effect of Studentization depended on the prediction dimension and the contamination pattern, with decreases in some pointwise comparisons and under severe contamination such as $r = 0.2$ or focused contamination of both inputs and responses.

Outside the assumptions. The comparison with existing robust GPs goes beyond the DPD analysis, covering learned kernel parameters and real data. Joint prediction improved in most tested conditions for each model, while pointwise improvements varied across models. The observed decreases were small, and Proposition 1 bounds the possible decrease when Studentization keeps the fitted mean and predictive scale fixed. These conclusions concern PLL; marginal coverage changed little in these experiments (Appendix G.4).

7 Conclusion

We asked when Studentization helps in GP regression with a shared variance scale, and when it does not. Because the Studentization gain at $\beta = 0$ is available exactly, the predictive cost of robustification can be quantified relative to this known finite-sample baseline.

Within the assumed model, Theorem 1 quantifies the cost of DPD fitting near $\beta = 0$ as a linear penalty in β plus a bounded quadratic remainder, and gives a sufficient condition for retaining a prescribed fraction of the exact gain. Controlled experiments supported this expansion and assessed retention over a finite grid of β values. Further, outside those assumptions, applying Studentization to three existing robust GP models with learned kernel parameters improved joint prediction in most tested conditions, with small decreases where it did not, and Proposition 1 bounds the possible decrease without assumptions on the data distribution or the fitting procedure.

The theoretical analysis is local in β and holds the kernel parameters fixed. In future work we aim to extend the analysis to settings with outliers and to relax the model assumptions, for example by considering Bayesian neural networks (Blundell et al., 2015).

AI USE STATEMENT

We used generative AI tools to assist with formulating and proving mathematical claims, implementation, running experiments and drafting and editing the manuscript. All the materials generated by AI tools were reviewed and edited by the authors. We take responsibility for the final manuscript, code, and reported results.

REFERENCES

J. Aitchison. Goodness of prediction fit. *Biometrika*, 62(3):547–554, 1975. doi: 10.1093/biomet/62.3.547. URL https://doi.org/10.1093/biomet/62.3.547.

Matías Altamirano, François-Xavier Briol, and Jeremias Knoblauch. Robust and conjugate gaussian process regression. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1155–1185. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/altamirano24a.html.

Sebastian Ament, Elizabeth Santorella, David Eriksson, Ben Letham, Maximilian Balandat, and Eytan Bakshy. Robust gaussian processes via relevance pursuit. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/7124771cf8c9b0f6e9c7fec0a66c5866-Abstract-Conference.html.

Arthur Asuncion and David J. Newman. UCI machine learning repository. University of California, Irvine, 2007. URL https://archive.ics.uci.edu.

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. Botorch: A framework for efficient monte-carlo bayesian optimization. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5b1b89d98b7286673128a5fb112cb9a-Abstract.html.

Ayanendranath Basu, Ian R Harris, Nils L Hjort, and Michael C Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 09 1998. ISSN 0006-3444. doi: 10.1093/biomet/85.3.549. URL https://doi.org/10.1093/biomet/85.3.549.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, 11 2016. ISSN 1369-7412. doi: 10.1111/rssb.12158. URL https://doi.org/10.1111/rssb.12158.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1613–1622. PMLR, 2015. URL https://proceedings.mlr.press/v37/blundell15.html.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16(5):1190–1208, 1995. doi: 10.1137/0916069. URL https://doi.org/10.1137/0916069.

DLMF. *NIST Digital Library of Mathematical Functions*. https://dlmf.nist.gov/, Release 1.2.8 of 2026-09-15. URL https://dlmf.nist.gov/. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

Rick Durrett. *Probability: theory and examples*. 5 edition, 2019. URL https://www.cambridge.org/core/books/probability/0A9D3B1F5C6E4A7B8D2F1C3E4D5F6A7B.

J. Gerritsma, R. Onnink, and A. Versluis. Yacht Hydrodynamics. UCI Machine Learning Repository, 1981. URL https://doi.org/10.24432/C5XG7R.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Mark S. Handcock and Michael L. Stein. A bayesian analysis of kriging. *Technometrics*, 35(4): 403–410, 1993. doi: 10.1080/00401706.1993.10485354. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1993.10485354.

David Harrison, Jr. and Daniel L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978. doi: 10.1016/0095-0696(78)90006-2. URL https://doi.org/10.1016/0095-0696(78)90006-2.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

Pasi Jylänki, Jarno Vanhatalo, and Aki Vehtari. Robust gaussian process regression with a student-*t* likelihood. *J. Mach. Learn. Res.*, 12:3227–3257, 2011. doi: 10.5555/1953048.2078209. URL https://dl.acm.org/doi/10.5555/1953048.2078209.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference. *J. Mach. Learn. Res.*, 23:132:1–132:109, 2022. URL https://jmlr.org/papers/v23/19-1047.html.

Feng Liang and Andrew R. Barron. Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Trans. Inf. Theory*, 50(11):2708–2726, 2004. doi: 10.1109/TIT.2004.836922. URL https://doi.org/10.1109/TIT.2004.836922.

Yifan Lu, Jiayi Ma, Leyuan Fang, Xin Tian, and Junjun Jiang. Robust and scalable gaussian process regression and its applications. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 21950–21959. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02102. URL https://doi.org/10.1109/CVPR52729.2023.02102.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006. ISBN 026218253X. URL https://www.worldcat.org/oclc/61285753.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 3 edition, 1976. URL https://search.worldcat.org/title/1502474.

Amar Shah, Andrew Gordon Wilson, and Zoubin Ghahramani. Student-t processes as alternatives to gaussian processes. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, volume 33 of *JMLR Workshop and Conference Proceedings*, pp. 877–885. JMLR.org, 2014. URL http://proceedings.mlr.press/v33/shah14.html.

I.M Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(67)90144-9. URL https://www.sciencedirect.com/science/article/pii/0041555367901449.

Qingtao Tang, Li Niu, Yisen Wang, Tao Dai, Wangpeng An, Jianfei Cai, and Shu-Tao Xia. Student-t process regression with student-t likelihood. In Carles Sierra (ed.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 2822–2828. ijcai.org, 2017. doi: 10.24963/IJCAI.2017/393. URL https://doi.org/10.24963/ijcai.2017/393.

Athanasios Tsanas and Angeliki Xifara. Energy Efficiency. UCI Machine Learning Repository, 2012. URL https://doi.org/10.24432/C51307.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018. doi: 10.1017/9781108231596. URL https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.html.

APPENDIX CONTENTS

A Tools for the proofs **14**
A.1 Existence of the fitted solution . . . 14
A.2 Smooth dependence on the robustification parameter . . . 14
A.3 Differentiation under the expectation . . . 14
A.4 Gaussian length and direction . . . 14
A.5 Cancellation by symmetry . . . 15
A.6 The quadratic remainder bound . . . 15
A.7 A beta variable from two gamma variables . . . 15

B Precision projection **15**
B.1 Proof sketch . . . 15
B.2 Proof . . . 16

C Proof of Lemma 1 **16**
C.1 Proof sketch . . . 16
C.2 Proof . . . 16

D Proof of Proposition 3 **17**
D.1 Proof sketch . . . 17
D.2 Proof . . . 17

E Proof of Theorem 1 **18**
E.1 Proof sketch . . . 18
E.2 Construction of the fitted solution . . . 18
E.3 Derivative bounds . . . 20
E.4 First-order predictive cost . . . 23
E.5 Remainder bound and gain retention . . . 26

F Experimental details and supplementary results **26**
F.1 Numerical implementation . . . 26
F.2 Finite-sample gain under the assumed GP model . . . 27
F.3 Prediction under contamination . . . 29

G Additional results for existing robust GPs **32**
G.1 Data and contamination . . . 32
G.2 Model fitting and prediction . . . 32
G.3 Variation across conditions and splits . . . 32
G.4 Predictive intervals . . . 37
G.5 Prediction blocks and the original data processing . . . 39
G.6 Optimization diagnostics . . . 39
G.7 Proof of Proposition 1 and numerical bounds . . . 39

A TOOLS FOR THE PROOFS

This section collects the standard mathematical results used in the proof of Theorem 1. The statements are adapted to our notation; the cited sources provide the original formulations, and we note any specific specializations below. Brown boxes visually distinguish these classical tools from the novel results established in this paper.

A.1 EXISTENCE OF THE FITTED SOLUTION

We use the contraction principle to prove that the iterative construction converges to a unique fixed point within the specified parameter cube.

Theorem A.1 (Banach contraction principle; Rudin, 1976, Theorem 9.23)**.** *On a nonempty complete metric space $(\mathcal{X}, d)$, suppose $T : \mathcal{X} \to \mathcal{X}$ satisfies $d(Tx, Ty) \le \kappa d(x, y)$ for all $x, y \in \mathcal{X}$, where $0 \le \kappa < 1$. There is a unique $x^\star$ with $Tx^\star = x^\star$, and $T^k x_0 \to x^\star$ for every $x_0 \in \mathcal{X}$.*

A.2 SMOOTH DEPENDENCE ON THE ROBUSTIFICATION PARAMETER

This theorem ensures that our constructed stationary solution depends smoothly on the robustification parameter β and provides the linear system for its derivatives. The following C^2 formulation follows from the C^1 theorem and its derivative formula in Rudin (1976, Theorem 9.28).

Theorem A.2 (Implicit function theorem)**.** *Let $F : \mathcal{U} \subset \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}^m$ be C^2 on an open set, with $F(x_0, t_0) = 0$ and $D_x F(x_0, t_0)$ invertible. Near (x_0, t_0), the equation $F(x, t) = 0$ has a unique solution $x = x(t)$, which is C^2 and satisfies*

$$Dx(t) = -\{D_x F(x(t), t)\}^{-1} D_t F(x(t), t).$$

A.3 DIFFERENTIATION UNDER THE EXPECTATION

To differentiate the predictive risk with respect to β, we need uniform integrable bounds that allow us to exchange differentiation and expectation. The following formulation combines differentiation under the integral sign with the dominated convergence theorem (Durrett, 2019, Theorems A.5.3 and 1.6.7).

Theorem A.3 (Dominated differentiation)**.** *Let $B > 0$ and let $F(b, \omega)$ be jointly measurable, with $b \mapsto F(b, \omega)$ twice continuously differentiable on $[0, B]$ for almost every ω. Suppose $\mathbb{E}|F(0, \omega)| < \infty$ and there are nonnegative random variables H_1, H_2, with $\mathbb{E}H_j < \infty$, such that*

$$\sup_{0 \le b \le B} |\partial_b^j F(b, \omega)| \le H_j(\omega), \qquad j = 1, 2,$$

outside a single event of probability zero. Then $b \mapsto \mathbb{E}[F(b, \omega)]$ is twice continuously differentiable and

$$\frac{d^j}{db^j} \mathbb{E}[F(b, \omega)] = \mathbb{E}[\partial_b^j F(b, \omega)], \qquad j = 1, 2.$$

Derivatives at 0 and B are one-sided.

A.4 GAUSSIAN LENGTH AND DIRECTION

Separating a Gaussian vector into its length and independent direction reduces high dimensional matrix expectations to scalar quantities. The distributional statement is Vershynin (2018, Exercise 3.3.7), and the second moment identities follow from rotational symmetry (Vershynin, 2018, Proposition 3.3.2).

Theorem A.4 (Gaussian radial decomposition)**.** *For $Z \sim \mathcal{N}(0, I_d)$, $d \ge 1$, define $R = \|Z\|_2$ and $U = Z/R$ on the event $R > 0$, which has probability one. Then $R^2 \sim \chi^2_d$, $U \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and R and U are independent. Moreover,*

$$\mathbb{E}[U] = 0, \qquad \mathbb{E}[UU^\top] = I_d/d, \qquad \mathbb{E}[ZZ^\top \mid R^2] = (R^2/d) I_d.$$

A.5 Cancellation by symmetry

Symmetry cancels odd terms in the derivative of the log predictive density, including the contribution from changes in the predictive mean. The following is a direct consequence of the change-of-variables formula for expectations in Durrett (2019, Theorem 1.6.9), applied to Z and $-Z$.

Theorem A.5 (Odd functions under a symmetric law). *Suppose $Z \in \mathbb{R}^d$ and $-Z$ have the same distribution. For any measurable $g : \mathbb{R}^d \to \mathbb{R}$ with $g(-z) = -g(z)$ and $\mathbb{E}|g(Z)| < \infty$, we have $\mathbb{E}[g(Z)] = 0$. The same conclusion holds under a conditional distribution that is symmetric, provided the conditional absolute expectation is finite.*

A.6 The quadratic remainder bound

This integral form of Taylor's theorem converts uniform bounds on second derivatives into a quadratic remainder bound for the local risk expansion. The identity follows from the fundamental theorem of calculus and integration by parts (Rudin, 1976, Theorems 6.21–6.22), applied to $(b-t)f''(t)$.

Theorem A.6 (Taylor's theorem with integral remainder). *Let $B > 0$ and let $f : [0, B] \to \mathbb{R}$ be twice continuously differentiable, with one-sided derivatives at the endpoints. For $0 \le b \le B$,*

$$f(b) - f(0) - bf'(0) = \int_0^b (b-t)f''(t)\,dt.$$

Consequently, $\sup_{0\le t\le B} |f''(t)| \le M$ implies $|f(b) - f(0) - bf'(0)| \le Mb^2/2$.

A.7 A beta variable from two gamma variables

This result separates a ratio from the sum, which simplifies the logarithmic expectation in the proof. It follows by changing variables in the gamma densities of Durrett (2019, Example 2.1.17 and the proof of Theorem 2.1.18); the independence follows from the density factorization below.

Theorem A.7 (Sum and ratio of independent gamma variables). *Let $X \sim \mathrm{Gam}(a, \lambda)$ and $Y \sim \mathrm{Gam}(b, \lambda)$ be independent, with $a, b, \lambda > 0$ and λ the common rate. Then $T = X + Y \sim \mathrm{Gam}(a + b, \lambda)$ and $B = X/(X + Y) \sim \mathrm{Beta}(a, b)$ are independent.*

The transformation $x = vt$, $y = (1 - v)t$ has Jacobian t, and the joint density factors as

$$f_{T,B}(t, v) = \frac{\lambda^{a+b} t^{a+b-1} e^{-\lambda t}}{\Gamma(a + b)} \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} v^{a-1}(1 - v)^{b-1},$$

for $t > 0$ and $0 < v < 1$.

B Precision projection

We state the projection result used in Section 4.2 and prove it by decomposing the KL objective over the class of distributions with the prescribed mean.

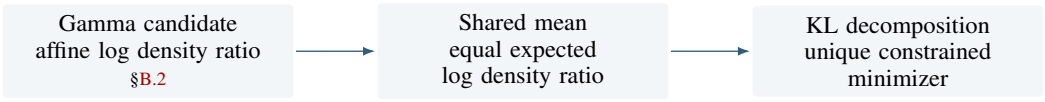
Proposition 2 (KL projection with a fixed mean precision). *Assume $Q_{\boldsymbol{y}} > 0$ and $0 < \widehat{\tau}^{(\beta)} < \infty$. Let $p_{\mathrm{ref}}(\tau \mid \boldsymbol{y}) = \mathrm{Gam}(\tau; n/2, Q_{\boldsymbol{y}}/2)$ be the precision posterior under the Gaussian model. Among probability densities q on $(0, \infty)$, the problem*

$$\min_q D_{\mathrm{KL}}(q(\tau) \,\|\, p_{\mathrm{ref}}(\tau \mid \boldsymbol{y})) \quad \textit{subject to} \quad \mathbb{E}_q[\tau] = \widehat{\tau}^{(\beta)} \tag{12}$$

has the unique solution $q_\beta^{(\tau)}(\tau \mid \mathbf{y} = \boldsymbol{y}) = \mathrm{Gam}(\tau; n/2, n/\{2\widehat{\tau}^{(\beta)}\})$.

B.1 Proof sketch

The log density ratio between the proposed Gamma candidate and the reference posterior is a constant plus a term linear in τ. Thus, all densities with the same mean $\widehat{\tau}^{(\beta)}$ have the same expected log density ratio. This permits a direct KL decomposition.



B.2 Proof

Fix observed responses with $Q_{\boldsymbol{y}} > 0$ and a fitted precision $0 < \widehat{\tau}^{(\beta)} < \infty$. Write $a = n/2$, $b = Q_{\boldsymbol{y}}/2$, and $c = a/\widehat{\tau}^{(\beta)}$. The candidate $q^\star(\tau) = \mathrm{Gam}(\tau; a, c)$ is a valid density with mean $\widehat{\tau}^{(\beta)}$. Its log density ratio is a constant plus a term linear in τ: $\log\{q^\star(\tau)/p_{\mathrm{ref}}(\tau \mid \boldsymbol{y})\} = a\log(c/b) + (b-c)\tau$. For any feasible density q satisfying the mean constraint, this form guarantees:

$$\mathbb{E}_q\left[\log \frac{q^\star(\tau)}{p_{\mathrm{ref}}(\tau \mid \boldsymbol{y})}\right] = a\log\frac{c}{b} + (b-c)\widehat{\tau}^{(\beta)} = \mathbb{E}_{q^\star}\left[\log \frac{q^\star(\tau)}{p_{\mathrm{ref}}(\tau \mid \boldsymbol{y})}\right].$$

For every such q with finite KL divergence to the reference, splitting the log ratio yields an exact KL decomposition:

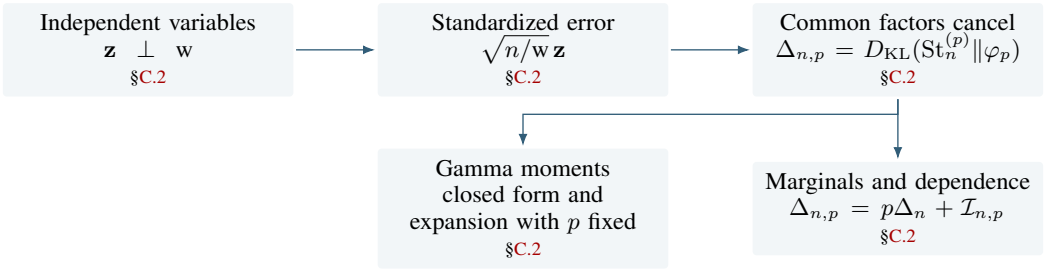
$$D_{\mathrm{KL}}(q\|p_{\mathrm{ref}}) = D_{\mathrm{KL}}(q\|q^\star) + D_{\mathrm{KL}}(q^\star\|p_{\mathrm{ref}}) \geq D_{\mathrm{KL}}(q^\star\|p_{\mathrm{ref}}). \tag{13}$$

Because $q^\star$ has finite KL divergence, densities with infinite objective value cannot be minimizers. The strict nonnegativity of KL divergence ensures $q^\star$ is the unique minimizer. At $\beta = 0$, $c = b$, so the projection recovers p_{ref}. □

C Proof of Lemma 1

C.1 Proof sketch

The standardized prediction error follows a Student-t distribution. The common determinant factors cancel in the log ratio, reducing the difference in predictive risk to the KL divergence between standard Student-t and Gaussian densities.



C.2 Proof

Gaussian conditioning yields two independent variables: the scaled training residual $\mathrm{w} = \tau_0 Q_{\mathbf{y}} \sim \chi^2(n)$ and the true standardized test error $\mathbf{z} = \sqrt{\tau_0}\boldsymbol{V}_*^{-1/2}(\mathbf{y}_* - \boldsymbol{\mu}_*^{(0)}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$, with $\mathrm{w} \perp \mathbf{z}$. Consequently, the empirically standardized error follows a true Student-t distribution under repeated sampling:

$$\mathbf{t} = (\boldsymbol{\Sigma}_*^{(0)})^{-1/2}(\mathbf{y}_* - \boldsymbol{\mu}_*^{(0)}) = \sqrt{n/\mathrm{w}}\,\mathbf{z} \sim \mathrm{St}_n^{(p)}(\mathbf{0}, \boldsymbol{I}_p). \tag{14}$$

Because the common determinant factors cancel in the predictive density ratio, the risk difference isolates the gain from Studentization:

$$\mathcal{R}_p(q_{\mathrm{G}}^{(p)}) - \mathcal{R}_p(q_0^{(p)}) = \mathbb{E}\left[\log \frac{\mathrm{St}_n^{(p)}(\mathbf{t})}{\varphi_p(\mathbf{t})}\right] = \Delta_{n,p} > 0. \tag{15}$$

To evaluate this KL divergence, let $\mathrm{u} = \|\mathbf{z}\|_2^2 \sim \chi^2(p)$, independent of w. Using Gamma distribution moments, we obtain $\mathbb{E}[\|\mathbf{t}\|_2^2] = pn/(n-2)$ and $\mathbb{E}[\log(1 + \mathrm{u}/\mathrm{w})] = \psi((n+p)/2) - \psi(n/2)$. Substituting these moments gives the closed form

$$\begin{aligned}\Delta_{n,p} = \log\frac{\Gamma((n+p)/2)}{\Gamma(n/2)} + \frac{p}{2}\log\frac{2}{n} + \frac{pn}{2(n-2)} \\ - \frac{n+p}{2}\left\{\psi\left(\frac{n+p}{2}\right) - \psi\left(\frac{n}{2}\right)\right\},\end{aligned} \tag{16}$$

where Γ and ψ are the gamma and digamma functions. Using their standard asymptotic expansions (DLMF, equations 5.11.1–5.11.2) gives $\Delta_{n,p} = p(p+6)/(4n^2) + O(n^{-3})$ for fixed p.

To separate the dependence term, we explicitly split the joint Student-t density from the product of its marginals. Write $t_j = [\mathbf{t}]_j$. Each component t_j has marginal density $\mathrm{St}_n^{(1)}$. Because $\varphi_p(\mathbf{t}) = \prod_{j=1}^p \varphi_1(t_j)$, the log density ratio decomposes as follows:

$$\Delta_{n,p} = \mathbb{E}_{\mathbf{t}}\left[\log \frac{\mathrm{St}_n^{(p)}(\mathbf{t})}{\prod_{j=1}^p \mathrm{St}_n^{(1)}(t_j)}\right] + \sum_{j=1}^p \mathbb{E}_{\mathbf{t}}\left[\log \frac{\mathrm{St}_n^{(1)}(t_j)}{\varphi_1(t_j)}\right] = \mathcal{I}_{n,p} + p\Delta_n. \tag{17}$$

This confirms that the joint gain is the sum of marginal gains ($p\Delta_n$) plus the dependence term ($\mathcal{I}_{n,p} \geq 0$). □

D Proof of Proposition 3

Proposition 3 (Optimality for a fixed density of standardized errors)**.** *Under the assumptions of Lemma 1, let h be a fixed probability density on $\mathbb{R}^p$, independent of the training responses, satisfying $\int_{\mathbb{R}^p} \mathrm{St}_n^{(p)}(\mathbf{t})|\log h(\mathbf{t})|\, d\mathbf{t} < \infty$. Define*

$$q_h^{(p)}(\mathbf{y}_* \mid \mathbf{y}) = |\boldsymbol{\Sigma}_*^{(0)}|^{-1/2} h\Big((\boldsymbol{\Sigma}_*^{(0)})^{-1/2}(\mathbf{y}_* - \boldsymbol{\mu}_*^{(0)})\Big). \tag{18}$$

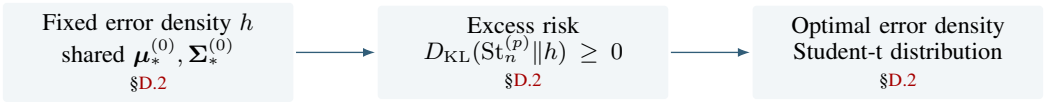
Then

$$\mathcal{R}_p(q_h^{(p)}) - \mathcal{R}_p(q_0^{(p)}) = D_{\mathrm{KL}}(\mathrm{St}_n^{(p)} \,\|\, h) \geq 0, \tag{19}$$

with equality if and only if $h = \mathrm{St}_n^{(p)}$ almost everywhere.

D.1 Proof sketch

As in Lemma 1, canceling the common factors in the density ratio leaves a KL divergence for any fixed density h of the standardized error, $D_{\mathrm{KL}}(\mathrm{St}_n^{(p)} \| h)$. Thus, the Student-t shape is the strictly optimal choice within this class.



D.2 Proof

The standardized error $\mathbf{t}$ in equation 14 has true density $\mathrm{St}_n^{(p)}$ under repeated sampling. Canceling the common determinant factor in the ratio of the Studentized predictor $q_0^{(p)}$ to the generic predictor $q_h^{(p)}$ gives:

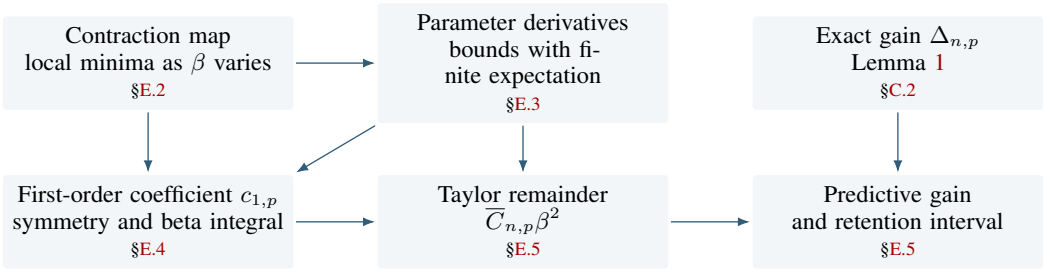
$$\mathcal{R}_p(q_h^{(p)}) - \mathcal{R}_p(q_0^{(p)}) = \mathbb{E}_{\mathbf{t}\sim \mathrm{St}_n^{(p)}}\left[\log \frac{\mathrm{St}_n^{(p)}(\mathbf{t})}{h(\mathbf{t})}\right] = D_{\mathrm{KL}}(\mathrm{St}_n^{(p)} \,\|\, h) \geq 0.$$

By the properties of KL divergence, equality holds if and only if $h = \mathrm{St}_n^{(p)}$ almost everywhere. □

E Proof of Theorem 1

E.1 Proof sketch

To calculate the predictive cost of robustification, we define the DPD fit as the fixed point of a contraction map. This guarantees a unique local solution that varies smoothly as β increases. We then uniformly bound the derivatives of these parameters to justify exchanging differentiation and expectation (Theorem A.3). Finally, we use the spherical symmetry of Gaussian errors to reduce matrix expectations to beta integrals and compute the exact first-order coefficient $c_{1,p}$.



E.2 Construction of the fitted solution

We use the solution constructed below to define the fitted predictor in Theorem 1. The theorem concerns this solution path from $\beta = 0$, rather than an arbitrary local minimum returned by an optimizer. For $\mathbf{y} \neq \mathbf{0}$, the Gaussian solution is

$$\boldsymbol{S}^{(0)} = (\boldsymbol{K}^{-1} + s^{-1}\boldsymbol{I}_n)^{-1}, \qquad \boldsymbol{m}^{(0)} = s^{-1}\boldsymbol{S}^{(0)}\mathbf{y}, \qquad \widehat{\tau}^{(0)} = n/Q_{\mathbf{y}}, \tag{20}$$

where $s = \sigma^2$. For generic $(\boldsymbol{m}, \boldsymbol{S}, \tau)$, the precision stationary equation uses

$$\Omega_\beta = \sum_i w_i - n\beta\xi_\beta, \qquad Q_\beta = \sum_i \frac{w_i e_i^2}{d_i} + \frac{\boldsymbol{m}^\top \boldsymbol{K}^{-1}\boldsymbol{m}}{\zeta_\beta(\tau)}. \tag{21}$$

Fix $\mathbf{y} \neq \mathbf{0}$ and write $s = \sigma^2$, $\kappa_i = K_{ii}/s$, and $\kappa = \max_i \kappa_i$, so $d_0 = n + 1 + \kappa$. Throughout the construction and derivative bounds, C denotes a finite positive universal upper bound, independent of $n, p, \boldsymbol{K}, s, \tau_0$, and $\mathbf{y}$. Its value may differ between occurrences: sums and products of such bounds are again denoted by C. We initially restrict $\beta d_0 \leq 1/4$ and give distinct names $C_{\mathrm{con}}, C_{\log}, C_{\mathrm{abs}}$ to the constants that further restrict this interval. They control contraction, the logarithmic bound on g, and absorption of the second-derivative term, respectively; Appendix E.3.2 collects these restrictions into one choice of c_0.

E.2.1 Stationary equations and fixed point map

The auxiliary variable $a_i = S_{ii}/K_{ii}$ is the ratio of fitted to prior marginal variance, while $b_i = \tau(y_i - m_i)^2/\{2n(s + \beta S_{ii})\}$ is a squared standardized residual divided by $2n$. These coordinates contain all the fitted parameter dependence of the observation weights apart from the common factor $g = \zeta_\beta(\tau)$. For fixed $(\beta, \mathbf{y})$ in the interval below, each input pair $(\boldsymbol{a}, \boldsymbol{b})$ determines a unique update $(\boldsymbol{m}, \boldsymbol{S}^+, \tau)$ by first forming the weights and then solving one scalar equation for g. The map returns updated coordinates $(\boldsymbol{a}^+, \boldsymbol{b}^+)$; at a fixed point, these coordinates agree with their definitions in terms of the fitted parameters. To construct their values, start with any $\boldsymbol{x} = (\boldsymbol{a}, \boldsymbol{b}) \in [0,1]^{2n}$ and set

$$\begin{gathered} w_i = t_i^{-1/2} e^{-n\beta b_i}, \quad \lambda_i = w_i/t_i, \quad c_i = 1 - 2n\beta b_i, \\ \boldsymbol{D} = s^{-1}\,\mathrm{diag}(\lambda_i), \qquad \text{and} \qquad \Omega = \sum_i w_i - n\beta(1+\beta)^{-3/2}, \end{gathered} \tag{22}$$

where $t_i = 1 + \beta\kappa_i a_i$ for $i = 1, \ldots, n$. Uniformly over the cube, the weights approach one as βd_0 decreases:

$$\max_i\{|1 - w_i|, |1 - \lambda_i|, |1 - c_i|\} + |1 - \Omega/n| \leq C\beta d_0. \tag{23}$$

On the preliminary range $\beta d_0 \leq 1/4$, we have $w_i \geq e^{-1/4}/\sqrt{5/4} > 1/2$, and $\lambda_i, c_i, \Omega/n$ are bounded away from zero by universal constants. For $g > 0$, define $\boldsymbol{K}_g = g\boldsymbol{K} + \boldsymbol{D}^{-1}$ and

$Q(g) = \mathbf{y}^\top \boldsymbol{K}_g^{-1}\mathbf{y}$. We abbreviate $Q = Q(g)$ at the selected root. Choose g as the solution of

$$g = \left(\frac{\Omega}{2\pi s\, Q(g)}\right)^{\beta/2}. \tag{24}$$

To see that this solution is unique, take logarithms of both sides of equation 24 and subtract the right side from the left. In the coordinate $r = \log g$, the resulting difference has derivative $1 - \beta\theta/2 > 0$, where $\theta = \mathbf{y}^\top \boldsymbol{K}_g^{-1} g\boldsymbol{K}\boldsymbol{K}_g^{-1}\mathbf{y}/Q \in [0, 1]$. This difference tends to $-\infty$ and $+\infty$ as $r \to -\infty$ and $r \to +\infty$, respectively, because $Q(e^r)$ tends to $\mathbf{y}^\top \boldsymbol{D}\mathbf{y}$ at the first limit and is asymptotic to $e^{-r}\mathbf{y}^\top \boldsymbol{K}^{-1}\mathbf{y}$ at the second. At $\beta = 0$, equation 24 gives $g = 1$.

With this g, define

$$\begin{gathered}\mathbf{e} = (\boldsymbol{I}_n + g\boldsymbol{K}\boldsymbol{D})^{-1}\mathbf{y}, \quad \boldsymbol{m} = \mathbf{y} - \mathbf{e}, \quad \tau = \Omega/Q, \\ \boldsymbol{S}^+ = (\boldsymbol{K}^{-1} + \boldsymbol{D}_c)^{-1}, \quad a_i^+ = \frac{S_{ii}^+}{K_{ii}}, \quad \text{and} \quad b_i^+ = \frac{\tau e_i^2}{2ns\, t_i},\end{gathered} \tag{25}$$

where $\boldsymbol{D}_c = g\boldsymbol{D}\,\mathrm{diag}(c_i)$. These equations define $\mathcal{T}_{\beta,\mathbf{y}}(\boldsymbol{x}) = (\boldsymbol{a}^+, \boldsymbol{b}^+)$. The denominator of b_i^+ uses the input coordinate a_i through t_i. The energy identity and the resulting bounds are

$$\begin{gathered}Q = \frac{\boldsymbol{m}^\top \boldsymbol{K}^{-1}\boldsymbol{m}}{g} + \mathbf{e}^\top \boldsymbol{D}\mathbf{e}, \\ \tau\mathbf{e}^\top \boldsymbol{D}\mathbf{e} \le \Omega \le n, \qquad \text{and} \qquad b_i^+ \le \frac{1}{2w_i} < 1.\end{gathered} \tag{26}$$

Together with $0 \prec \boldsymbol{S}^+ \preceq \boldsymbol{K}$, these bounds show that the map sends the cube into itself.

E.2.2 Contraction and smoothness

For an infinitesimal change $\mathrm{d}\boldsymbol{x}$ at fixed $\beta, \mathbf{y}$, write $r_x = d_0\beta\|\mathrm{d}\boldsymbol{x}\|_\infty$. Differentiating equation 22 gives $|\mathrm{d}\log\Omega| + \|\mathrm{d}\log\boldsymbol{\lambda}\|_\infty \le Cr_x$. Differentiating the inverse matrix in Q gives

$$\mathrm{d}\log Q = -\theta\,\mathrm{d}\log g + \varepsilon_Q, \qquad \text{with} \qquad |\varepsilon_Q| \le (1-\theta)\|\mathrm{d}\log\boldsymbol{\lambda}\|_\infty.$$

Here ε_Q is the contribution to $\mathrm{d}\log Q$ from changes in $\boldsymbol{\lambda}$ at fixed g. Equation 24 therefore implies

$$\begin{gathered}\mathrm{d}\log g = \frac{\beta}{2-\beta\theta}(\mathrm{d}\log\Omega - \varepsilon_Q), \\ |\mathrm{d}\log g| \le C\beta r_x, \qquad \text{and} \qquad |\mathrm{d}\log\tau| \le Cr_x.\end{gathered}$$

The residual derivative satisfies

$$\begin{gathered}\boldsymbol{D}^{1/2}\mathrm{d}\mathbf{e} = -\boldsymbol{F}\{\mathrm{d}\log g\,\boldsymbol{I}_n + \mathrm{diag}(\mathrm{d}\log\boldsymbol{\lambda})\}\boldsymbol{D}^{1/2}\mathbf{e}, \\ \text{where} \quad \boldsymbol{F} = (\boldsymbol{I}_n + g\boldsymbol{D}^{1/2}\boldsymbol{K}\boldsymbol{D}^{1/2})^{-1}g\boldsymbol{D}^{1/2}\boldsymbol{K}\boldsymbol{D}^{1/2}.\end{gathered}$$

Because $0 \preceq \boldsymbol{F} \preceq \boldsymbol{I}_n$, its operator norm is at most one. Write $\mathrm{d}\omega_i = \mathrm{d}\log g + \mathrm{d}\log\lambda_i + \mathrm{d}\log c_i$ and $\mathrm{d}u_i = \mathrm{d}\log t_i$. Their maximum norms are bounded by Cr_x. The identity $\mathrm{d}\boldsymbol{S}^+ = -\boldsymbol{S}^+\boldsymbol{D}_c\,\mathrm{diag}(\mathrm{d}\boldsymbol{\omega})\boldsymbol{S}^+$ and the order $\boldsymbol{S}^+\boldsymbol{D}_c\boldsymbol{S}^+ \preceq \boldsymbol{S}^+ \preceq \boldsymbol{K}$ give

$$|\mathrm{d}a_i^+| \le \|\mathrm{d}\boldsymbol{\omega}\|_\infty \frac{[\boldsymbol{S}^+\boldsymbol{D}_c\boldsymbol{S}^+]_{ii}}{K_{ii}} \le Cr_x.$$

For the residual coordinate, differentiation and equation 26 give

$$\begin{gathered}\mathrm{d}b_i^+ = b_i^+(\mathrm{d}\log\tau - \mathrm{d}u_i) + \frac{\tau e_i\,\mathrm{d}e_i}{nst_i}, \\ \left|\frac{\tau e_i\,\mathrm{d}e_i}{nst_i}\right| \le \frac{\tau}{nw_i}\|\boldsymbol{D}^{1/2}\mathbf{e}\|_2\|\boldsymbol{D}^{1/2}\mathrm{d}\mathbf{e}\|_2 \le Cr_x.\end{gathered}$$

Thus $\max_i\{|\mathrm{d}a_i^+|, |\mathrm{d}b_i^+|\} \le Cr_x$, including when $e_i = 0$. Integrating along line segments in the cube proves

$$\|\mathcal{T}_{\beta,\mathbf{y}}(\boldsymbol{x}) - \mathcal{T}_{\beta,\mathbf{y}}(\widetilde{\boldsymbol{x}})\|_\infty \le C_{\mathrm{con}}d_0\beta\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_\infty \tag{27}$$

for all $\boldsymbol{x}, \widetilde{\boldsymbol{x}} \in [0, 1]^{2n}$. Impose $C_{\mathrm{con}}c_0 \le 1/2$, so the contraction factor is at most $1/2$ on $\beta d_0 \le c_0$. Theorem A.1, applied to the closed cube with the maximum-norm metric, then gives a unique fixed point there. The iteration in Theorem A.1, started at a fixed initial value in the cube such as $\mathbf{0}$,

has continuous iterates in nonzero $\mathbf{y}$ and hence a measurable limit. The nonzero derivative of the scalar log equation permits Theorem A.2 to be applied to its root, so the root and the fixed point map are twice continuously differentiable near each fixed point. Also, $\boldsymbol{I}_{2n} - D_{\boldsymbol{x}}\mathcal{T}_{\beta,\mathbf{y}}$ is invertible by equation 27. Applying Theorem A.2 to $F(\boldsymbol{x}, \beta) = \boldsymbol{x} - \mathcal{T}_{\beta,\mathbf{y}}(\boldsymbol{x})$ gives a C^2 solution path, with one-sided derivatives at the endpoints. At $\beta = 0$, equations 22, 24, and 25 recover equation 20.

E.2.3 Stationarity and local minimality

For $U = (\boldsymbol{m}, \boldsymbol{S}, \tau)$ and fixed $\boldsymbol{x}$, define the stationary residuals

$$\begin{aligned} R_1 &= \boldsymbol{K}^{-1}\boldsymbol{m} - g\boldsymbol{D}\mathbf{e}, \\ R_2 &= \boldsymbol{S}^{-1} - \boldsymbol{K}^{-1} - g\boldsymbol{D}\,\mathrm{diag}(1 - 2n\beta b_i), \\ \text{and} \quad R_3 &= \tau E - \Omega, \end{aligned}$$

where $E = \boldsymbol{m}^\top \boldsymbol{K}^{-1}\boldsymbol{m}/g + \mathbf{e}^\top \boldsymbol{D}\mathbf{e}$, $g = (\tau/(2\pi s))^{\beta/2}$, and $\mathbf{e} = \mathbf{y} - \boldsymbol{m}$. Equations 24 and 25 solve $R = (R_1, R_2, R_3) = 0$ for fixed $\boldsymbol{x}$. Define $H_{a,i} = S_{ii}/K_{ii}$ and $H_{b,i} = \tau e_i^2/\{2ns(1 + \beta\kappa_i a_i)\}$. Imposing the auxiliary equations $\boldsymbol{x} = H(U, \boldsymbol{x})$ recovers the weights of the original objective. Its gradients then satisfy

$$\nabla_{\boldsymbol{m}} J_\beta = \tau R_1, \qquad \nabla_{\boldsymbol{S}} J_\beta = -R_2/2, \qquad \text{and} \qquad \partial_\tau J_\beta = gR_3/(2\tau). \tag{28}$$

Thus the fixed point is an interior stationary point. At $\beta = 0$, we use the smooth extension of J_β obtained in Appendix E.4.2.

To establish local minimality, first note that R_U is invertible. Indeed, $\nabla_{\boldsymbol{m}} E = 2R_1/g = 0$ at the solution, and ordering the variables as $(\tau, \boldsymbol{m}, \boldsymbol{S})$ and residuals as (R_3, R_1, R_2) makes R_U block lower triangular. Its diagonal blocks are

$$Q(1 - \beta\theta/2) > 0, \qquad \boldsymbol{K}^{-1} + g\boldsymbol{D} \succ 0, \qquad \text{and} \qquad \delta\boldsymbol{S} \longmapsto -\boldsymbol{S}^{-1}\delta\boldsymbol{S}\boldsymbol{S}^{-1}.$$

The last operator is invertible on symmetric matrices. Also, $H_{\boldsymbol{x}}$ has only a lower-left block, so $\boldsymbol{I}_{2n} - H_{\boldsymbol{x}}$ is invertible. The Jacobian of the expanded equations is

$$\begin{pmatrix} R_U & R_{\boldsymbol{x}} \\ -H_U & \boldsymbol{I}_{2n} - H_{\boldsymbol{x}} \end{pmatrix}.$$

Eliminating U gives the Schur complement $\boldsymbol{I}_{2n} - D_{\boldsymbol{x}}\mathcal{T}$, which is invertible by equation 27. Eliminating $\boldsymbol{x}$ instead gives $R_U + R_{\boldsymbol{x}}(\boldsymbol{I}_{2n} - H_{\boldsymbol{x}})^{-1}H_U$, the Jacobian of the original stationary equations. This Jacobian is therefore invertible, and equation 28 makes the objective Hessian nondegenerate. At $\beta = 0$, its quadratic form is

$$\begin{aligned} \delta^2 J_0 = \tau\,\delta\boldsymbol{m}^\top(\boldsymbol{K}^{-1} + s^{-1}\boldsymbol{I}_n)\delta\boldsymbol{m} \\ + \frac{1}{2}\,\mathrm{Tr}(\boldsymbol{S}^{-1}\delta\boldsymbol{S}\boldsymbol{S}^{-1}\delta\boldsymbol{S}) + \frac{n}{2\tau^2}(\delta\tau)^2 > 0 \end{aligned} \tag{29}$$

for every nonzero perturbation. The Hessian is symmetric and continuous along the solution path. Because the Hessian is initially positive definite, an eigenvalue could become nonpositive only by passing through zero. This would contradict the nondegeneracy just established. This proves strict local minimality throughout $[0, b_0]$; invertibility of the fixed point Jacobian alone would not be enough. Finally, write $\boldsymbol{V}_*^{(\beta)} = \boldsymbol{R}_* + \boldsymbol{A}_*\boldsymbol{S}^{(\beta)}\boldsymbol{A}_*^\top$. Then $\boldsymbol{R}_* \succeq 0$ and $\boldsymbol{V}_*^{(0)} \succ 0$ imply $\boldsymbol{V}_*^{(\beta)} \succ 0$: a null vector of $\boldsymbol{R}_* + \boldsymbol{A}_*\boldsymbol{S}^{(\beta)}\boldsymbol{A}_*^\top$ would also be a null vector at $\beta = 0$.

E.3 Derivative bounds

Write $\partial_\beta = \partial/\partial\beta$ and $\partial_\beta^2 = \partial^2/\partial\beta^2$. We use right derivatives at zero. All derivatives in this subsection are taken along the solution path at fixed $\mathbf{y}$ unless stated otherwise. The norm $\|\cdot\|_\infty$ is the maximum norm for vectors and its induced matrix norm; $\|\cdot\|_2$ is the Euclidean norm for vectors and the operator norm for matrices. Define $L_{\mathbf{y}} = 1 + |\log\{n/(2\pi s Q_{\mathbf{y}})\}|$. For $j = 1, 2$, let $X_j = \|\partial_\beta^j \boldsymbol{x}\|_\infty$ denote the norms of the derivatives of the fixed point. We abbreviate $L = L_{\mathbf{y}}$ below.

E.3.1 Derivatives of the fixed point

The training responses enter the derivative bounds through the logarithmic factor L. Equation 22 gives $|\log(\Omega/n)| + \|\log\boldsymbol{\lambda}\|_\infty \le C\beta d_0$. Comparison of $\boldsymbol{K}_g = g\boldsymbol{K} + \boldsymbol{D}^{-1}$ with $\boldsymbol{K} + s\boldsymbol{I}_n$ in the

positive semidefinite order gives $|\log(Q/Q_{\mathbf{y}})| \le |\log g| + C\beta d_0$. Substitution in equation 24 gives the following bound, whose last inequality holds when $C_{\log} c_0 \le 1/2$:

$$|\log g| \le \frac{\beta}{2-\beta}\left\{\left|\log \frac{n}{2\pi s Q_{\mathbf{y}}}\right| + C_{\log}\beta d_0\right\} \le \beta L. \tag{30}$$

With $\boldsymbol{x}$ held fixed, the derivatives of $\log\Omega$, $\log\boldsymbol{\lambda}$, and $\log \boldsymbol{c}$ are bounded by Cd_0. Differentiating the scalar root then gives $|\partial_\beta \log g| \le CL$. The covariance and residual identities in Appendix E.2.2, together with equation 26, therefore imply

$$\|\partial_\beta \mathcal{T}_{\beta,\mathbf{y}}\|_\infty \le Cd_0 L, \qquad \text{and} \qquad X_1 \le Cd_0 L. \tag{31}$$

The second bound follows from the derivative formula in Theorem A.2, applied to $\boldsymbol{x} = \mathcal{T}_{\beta,\mathbf{y}}(\boldsymbol{x})$, and the uniformly bounded inverse of $\boldsymbol{I}_{2n} - D_{\boldsymbol{x}}\mathcal{T}$.

For the second derivative, write $u_i = \log t_i$, $v_i = \log\lambda_i$, and $z_i = \log c_i$. Also set $o = \log\Omega$, $r = \log g$, and $h = \log\tau$. The definitions and the scalar root imply

$$v_i = -3u_i/2 - n\beta b_i, \quad h = o - \log Q, \quad \text{and} \quad r = \frac{\beta}{2}\{h - \log(2\pi s)\}.$$

Differentiating the inverse matrix gives

$$\begin{aligned} |\partial_\beta \log Q| &\le |\partial_\beta r| + \|\partial_\beta \boldsymbol{v}\|_\infty, \\ \text{and} \quad |\partial_\beta^2 \log Q| &\le |\partial_\beta^2 r| + \|\partial_\beta^2 \boldsymbol{v}\|_\infty + C\big(|\partial_\beta r| + \|\partial_\beta \boldsymbol{v}\|_\infty\big)^2. \end{aligned} \tag{32}$$

To obtain the second inequality, decompose the derivatives of $\boldsymbol{K}_g$ into the positive semidefinite components $g\boldsymbol{K}$ and $\boldsymbol{D}^{-1}$. Their sum, normalized by $\boldsymbol{K}_g^{-1/2}$, is the identity, so the componentwise log derivatives bound the corresponding operator norms.

Equation 31 bounds the first derivatives of $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{z}, o, h$ by Cd_0L and that of r by CL. For $i = 1, \dots, n$, the second derivatives satisfy

$$\begin{aligned} \partial_\beta^2 u_i &= \frac{\kappa_i(2\partial_\beta a_i + \beta\partial_\beta^2 a_i)}{1+\beta\kappa_i a_i} - (\partial_\beta u_i)^2, \\ \partial_\beta^2 v_i &= -3\partial_\beta^2 u_i/2 - n(2\partial_\beta b_i + \beta\partial_\beta^2 b_i), \\ \text{and} \quad \partial_\beta^2 z_i &= -\frac{2n(2\partial_\beta b_i + \beta\partial_\beta^2 b_i)}{1-2n\beta b_i} - (\partial_\beta z_i)^2. \end{aligned}$$

The scalar identity for r also gives $\partial_\beta^2 r = \partial_\beta h + \beta\partial_\beta^2 h/2$. The first two derivatives of $f(\beta) = \beta(1+\beta)^{-3/2}$ are uniformly bounded on the chosen interval. Differentiating $\Omega = \sum_i w_i - nf(\beta)$ therefore bounds $\partial_\beta^2 o$ by $Cd_0^2L^2 + C\beta d_0 X_2$, as for $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{z}$. Substituting these bounds into equation 32 and absorbing the term $\beta|\partial_\beta^2 h|/2$ gives the same bound for h and r. Consequently,

$$\begin{aligned} \|\partial_\beta \boldsymbol{\omega}\|_\infty &\le Cd_0 L, \\ \text{and} \quad \|\partial_\beta^2 \boldsymbol{\omega}\|_\infty &\le Cd_0^2 L^2 + C\beta d_0 X_2, \end{aligned}$$

where $\omega_i = \log([\boldsymbol{D}_c]_{ii}) = r + v_i + z_i - \log s$.

We now close the bound on X_2. Differentiating $\boldsymbol{S} = (\boldsymbol{K}^{-1} + \boldsymbol{D}_c)^{-1}$ and using $\boldsymbol{S}\boldsymbol{D}_c\boldsymbol{S} \preceq \boldsymbol{S}$ gives

$$\begin{aligned} \|\boldsymbol{S}^{-1/2}(\partial_\beta \boldsymbol{S})\boldsymbol{S}^{-1/2}\|_2 &\le \|\partial_\beta \boldsymbol{\omega}\|_\infty, \\ \text{and} \quad \|\boldsymbol{S}^{-1/2}(\partial_\beta^2 \boldsymbol{S})\boldsymbol{S}^{-1/2}\|_2 &\le C\|\partial_\beta \boldsymbol{\omega}\|_\infty^2 + \|\partial_\beta^2 \boldsymbol{\omega}\|_\infty. \end{aligned}$$

Because $\boldsymbol{S} \preceq \boldsymbol{K}$, these inequalities bound $\|\partial_\beta^2 \boldsymbol{a}\|_\infty$ by $Cd_0^2L^2 + C\beta d_0 X_2$.

For the residual, set $j_i = r + v_i$ and $\boldsymbol{F}_e = (\boldsymbol{I}_n + g\boldsymbol{K}\boldsymbol{D})^{-1} g\boldsymbol{K}\boldsymbol{D}$. Differentiating $(\boldsymbol{I}_n + g\boldsymbol{K}\boldsymbol{D})\mathbf{e} = \mathbf{y}$ gives

$$\begin{aligned} \partial_\beta \mathbf{e} &= -\boldsymbol{F}_e \operatorname{diag}(\partial_\beta \boldsymbol{j})\mathbf{e}, \\ \text{and} \quad \partial_\beta^2 \mathbf{e} &= -\boldsymbol{F}_e\big\{\operatorname{diag}(\partial_\beta^2 \boldsymbol{j} + (\partial_\beta \boldsymbol{j})^2)\mathbf{e} \\ &\qquad + 2\operatorname{diag}(\partial_\beta \boldsymbol{j})\partial_\beta \mathbf{e}\big\}, \end{aligned}$$

where $(\partial_\beta \boldsymbol{j})^2$ is componentwise. The $\boldsymbol{D}$-weighted operator norm of $\boldsymbol{F}_e$ is at most one, which gives

$$E_1 := \frac{\|\boldsymbol{D}^{1/2}\partial_\beta \mathbf{e}\|_2}{\|\boldsymbol{D}^{1/2}\mathbf{e}\|_2} \le Cd_0L,$$
$$\text{and} \quad E_2 := \frac{\|\boldsymbol{D}^{1/2}\partial_\beta^2 \mathbf{e}\|_2}{\|\boldsymbol{D}^{1/2}\mathbf{e}\|_2} \le Cd_0^2L^2 + C\beta d_0 X_2.$$

Set $\alpha_i = h - u_i$, so $b_i = e^{\alpha_i} e_i^2/(2ns)$. Differentiating this expression twice and using equation 26 yields

$$\begin{aligned}|\partial_\beta^2 b_i| &\le \frac{\|\partial_\beta^2 \boldsymbol{\alpha}\|_\infty + \|\partial_\beta \boldsymbol{\alpha}\|_\infty^2}{2w_i} + \frac{2\|\partial_\beta \boldsymbol{\alpha}\|_\infty E_1 + E_1^2 + E_2}{w_i} \\ &\le Cd_0^2L^2 + C\beta d_0 X_2.\end{aligned}$$

This calculation remains valid when an individual residual is zero. Combining the bounds for $\boldsymbol{a}$ and $\boldsymbol{b}$ gives an inequality with X_2 on both sides. Impose $C_{\text{abs}}c_0 \le 1/2$ so that moving the last term to the left leaves at least $X_2/2$:

$$X_2 \le Cd_0^2L^2 + C_{\text{abs}}\beta d_0 X_2, \qquad \text{hence} \qquad X_2 \le 2Cd_0^2L^2. \tag{33}$$

E.3.2 Propagation to the predictive density

Write $\boldsymbol{V}_*^{(\beta)} = \boldsymbol{R}_* + \boldsymbol{A}_* \boldsymbol{S}^{(\beta)} \boldsymbol{A}_*^\top$, so $\boldsymbol{V}_*^{(0)} = \boldsymbol{V}_*$. In this subsection, abbreviate $\boldsymbol{\mu} = \boldsymbol{\mu}_*^{(\beta)}$, $\boldsymbol{V} = \boldsymbol{V}_*^{(\beta)}$, and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_*^{(\beta)}$. The preceding bounds imply

$$\|\boldsymbol{S}^{-1/2}(\partial_\beta^j \boldsymbol{S})\boldsymbol{S}^{-1/2}\|_2 + |\partial_\beta^j h| \le C(d_0L)^j \quad \text{for } j = 1, 2.$$

The order $\boldsymbol{A}_* \boldsymbol{S} \boldsymbol{A}_*^\top \preceq \boldsymbol{V}$ transfers the relative covariance bounds to $\boldsymbol{V}$ and then to $\boldsymbol{\Sigma} = e^{-h}\boldsymbol{V}$.

For the mean, let $\boldsymbol{C}_m = (\boldsymbol{K}^{-1} + g\boldsymbol{D})^{-1}$. Because $\boldsymbol{S}^{-1} \preceq \boldsymbol{C}_m^{-1}$, differentiating the mean equation and using equation 26 gives

$$\sqrt{\tau}\|\boldsymbol{S}^{-1/2}\partial_\beta \boldsymbol{m}\|_2 \le \sqrt{ng}\|\partial_\beta \boldsymbol{j}\|_\infty,$$
$$\text{and} \quad \sqrt{\tau}\|\boldsymbol{S}^{-1/2}\partial_\beta^2 \boldsymbol{m}\|_2 \le C\sqrt{ng}\{\|\partial_\beta^2 \boldsymbol{j}\|_\infty + \|\partial_\beta \boldsymbol{j}\|_\infty^2\}.$$

For $j = 1, 2$, applying $\boldsymbol{A}_*$ and using the same matrix order gives

$$\begin{aligned}A_j &:= \|\boldsymbol{\Sigma}^{-1/2}\partial_\beta^j \boldsymbol{\mu}\|_2 \le C\sqrt{ng}(d_0L)^j, \\ \text{and} \quad B_j &:= \|\boldsymbol{\Sigma}^{-1/2}(\partial_\beta^j \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1/2}\|_2 \le C(d_0L)^j.\end{aligned} \tag{34}$$

Let $\ell_\beta = \log q_\beta^{(p)}(\mathbf{y}_* \mid \mathbf{y})$ and let $\delta_* = (\mathbf{y}_* - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}_* - \boldsymbol{\mu})$ be the squared Mahalanobis distance. Use primes for derivatives in β in the following calculation and write $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$. Differentiating the quadratic form, including both its mean and scale, gives

$$|\delta_*'| \le 2A_1\sqrt{\delta_*} + B_1\delta_*,$$
$$|\delta_*''| \le 2A_1^2 + (2A_2 + 4A_1B_1)\sqrt{\delta_*} + (2B_1^2 + B_2)\delta_*.$$

The log density derivatives are

$$\ell_\beta' = -\tfrac{1}{2}\operatorname{Tr}(\boldsymbol{W}\boldsymbol{\Sigma}') - \tfrac{n+p}{2}\frac{\delta_*'}{n+\delta_*},$$
$$\ell_\beta'' = -\tfrac{1}{2}\operatorname{Tr}(\boldsymbol{W}\boldsymbol{\Sigma}'' - \boldsymbol{W}\boldsymbol{\Sigma}'\boldsymbol{W}\boldsymbol{\Sigma}') - \tfrac{n+p}{2}\left\{\frac{\delta_*''}{n+\delta_*} - \frac{(\delta_*')^2}{(n+\delta_*)^2}\right\}.$$

The trace terms are bounded by pB_1 and $p(B_2 + B_1^2)$, respectively. Using these bounds and $\delta_*/(n+\delta_*) \le 1$ and $2\sqrt{\delta_*}/(n+\delta_*) \le 1/\sqrt{n}$ gives, uniformly in $\mathbf{y}_*$,

$$|\partial_\beta \ell_\beta| \le C(n+p)(A_1/\sqrt{n} + B_1),$$
$$\text{and} \quad |\partial_\beta^2 \ell_\beta| \le C(n+p)(A_1^2/n + A_2/\sqrt{n} + A_1B_1/\sqrt{n} + B_1^2 + B_2).$$

Equations 30 and 34 therefore give a universal C such that

$$\begin{aligned} \sup_{0\le\beta\le b_0} |\partial_\beta \ell_\beta| &\le H_{1,p}(\mathbf{y}) := C(n+p)d_0 L_{\mathbf{y}}(1+e^{b_0 L_{\mathbf{y}}/2}), \\ \text{and} \quad \sup_{0\le\beta\le b_0} |\partial_\beta^2 \ell_\beta| &\le H_{2,p}(\mathbf{y}) := C(n+p)d_0^2 L_{\mathbf{y}}^2(1+e^{b_0 L_{\mathbf{y}}}). \end{aligned} \tag{35}$$

The restrictions introduced above are all satisfied by the single universal choice

$$c_0 = \min\left\{\frac{1}{4}, \frac{1}{2C_{\text{con}}}, \frac{1}{2C_{\log}}, \frac{1}{2C_{\text{abs}}}\right\}, \qquad b_0 = \frac{c_0}{d_0}.$$

The three constants are finite on the respective preliminary ranges used above and independent of the training responses. This gives a sufficient cutoff; their numerical values, and hence the most restrictive threshold, are not evaluated here. Because $n > 2$ and $d_0 \ge n+1$, this choice also gives $b_0 \le 1/\{4(n+1)\} < n/2$.

Under the assumed GP model, $\mathrm{w} = \tau_0 Q_{\mathbf{y}} \sim \chi^2_n$ and $L_{\mathbf{y}} = 1 + |a_0 - \log \mathrm{w}|$, where $a_0 = \log(n\tau_0/(2\pi s))$. The inequality $b_0 < n/2$ ensures that the upper bounds $H_{1,p}(\mathbf{y})$ and $H_{2,p}(\mathbf{y})$ have finite expectations. Near zero the relevant integrand is bounded by a constant times $\mathrm{w}^{n/2-1-b_0}(1+|\log \mathrm{w}|^2)$; at infinity the chi-square density dominates the polynomial and logarithmic factors. The baseline log density ℓ_0 is integrable because $n > 2$ and $\mathrm{w} \sim \chi^2_n$. Theorem A.3, with $F = \ell_\beta$ and the bounds in equation 35, therefore justifies differentiation under the expectation on $[0, b_0]$. We may define the predictor arbitrarily at $\mathbf{y} = \mathbf{0}$, which has probability zero under the assumed GP model and does not affect the risk.

E.4 First-order predictive cost

We calculate the derivative of predictive risk with respect to β at $\beta = 0$, using the derivatives of the fitted parameters and rotational symmetry.

E.4.1 Gaussian averaging of the DPD objective

For $p(z \mid f_i, \tau) = \mathcal{N}(z; f_i, \sigma^2/\tau)$, the integral term in the DPD loss is independent of f_i:

$$\frac{1}{1+\beta}\int p(z \mid f_i, \tau)^{1+\beta}\, dz = \left(\frac{\tau}{2\pi\sigma^2}\right)^{\beta/2}(1+\beta)^{-3/2} = \zeta_\beta(\tau)\xi_\beta.$$

For the term evaluated at the observed response y_i, write $\mathrm{f}_i = m_i + \sqrt{v_i/\tau}\, Z$ with $Z \sim \mathcal{N}(0,1)$ under q, and set $A = \beta v_i/\sigma^2$ and $B = \beta e_i\sqrt{\tau v_i}/\sigma^2$. Completing the square in the standard Gaussian integral gives

$$\begin{aligned} \mathbb{E}[e^{-AZ^2/2+BZ}] &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-(1+A)z^2/2+Bz}\, dz \\ &= (1+A)^{-1/2}\exp\left\{\frac{B^2}{2(1+A)}\right\}. \end{aligned}$$

Consequently,

$$\begin{aligned} \mathbb{E}_q[p(y_i \mid \mathrm{f}_i, \tau)^\beta] &= \zeta_\beta(\tau)e^{-\beta\tau e_i^2/(2\sigma^2)}\mathbb{E}[e^{-AZ^2/2+BZ}] \\ &= \zeta_\beta(\tau)\sqrt{\frac{\sigma^2}{\sigma^2+\beta v_i}}\exp\left\{-\frac{\beta\tau e_i^2}{2(\sigma^2+\beta v_i)}\right\} = \zeta_\beta(\tau)w_i. \end{aligned}$$

Thus $d_i = \sigma^2 + \beta v_i$ accounts for the variational uncertainty in f_i when averaging the density power. Substitution into equation 5, summation over observations, and addition of the Gaussian KL penalty give equation 7, with

$$D_{\mathrm{KL}}(q(\mathbf{f} \mid \tau)\|p(\mathbf{f} \mid \tau)) = \frac{1}{2}\left\{\mathrm{Tr}(\boldsymbol{K}^{-1}\boldsymbol{S}) + \tau\boldsymbol{m}^\top\boldsymbol{K}^{-1}\boldsymbol{m} - n + \log\frac{|\boldsymbol{K}|}{|\boldsymbol{S}|}\right\}.$$

E.4.2 Expansion of the conditional objective

The first-order term of the DPD objective determines how the stationary solution changes with β. For generic $(\boldsymbol{m}, \boldsymbol{S}, \tau)$, write $e_i = \mathrm{y}_i - m_i$ and $v_i = [\boldsymbol{S}]_{ii}$. Define $\delta_i = \tau e_i^2/\sigma^2$, $\upsilon_i = v_i/\sigma^2$, and

$\ell = \frac{1}{2}\log\{\tau/(2\pi\sigma^2)\}$. The DPD loss has the local expansion

$$S_\beta(p, y) = -\log p(y) + \beta\left\{\int p(z)\log p(z)\,dz - 1 - \frac{1}{2}\log^2 p(y)\right\} + O(\beta^2).$$

The Gaussian observation density is $p(z \mid f_i, \tau) = \mathcal{N}(z; f_i, \sigma^2/\tau)$. The required moments of its log density satisfy

$$\int p(z \mid f_i, \tau)\log p(z \mid f_i, \tau)\,dz = \ell - \frac{1}{2},$$

$$\text{and} \quad \mathbb{E}_q[\log^2 p(y_i \mid \mathrm{f}_i, \tau)] = \ell^2 - \ell(\delta_i + \upsilon_i) + \frac{1}{4}(\delta_i^2 + 6\delta_i\upsilon_i + 3\upsilon_i^2).$$

Thus $J_\beta = J_0 + \beta J_1 + O(\beta^2)$ locally, where

$$\begin{aligned} J_0 &= \frac{n}{2}\log\frac{2\pi\sigma^2}{\tau} + \frac{\tau}{2\sigma^2}\|\mathbf{y} - \boldsymbol{m}\|_2^2 + \frac{1}{2\sigma^2}\operatorname{Tr}(\boldsymbol{S}) \\ &\quad + \frac{1}{2}\left\{\operatorname{Tr}(\boldsymbol{K}^{-1}\boldsymbol{S}) + \log\frac{|\boldsymbol{K}|}{|\boldsymbol{S}|} - n + \tau\boldsymbol{m}^\top\boldsymbol{K}^{-1}\boldsymbol{m}\right\}, \\ J_1 &= \sum_{i=1}^n\left\{\ell - \frac{3}{2} - \frac{\ell^2}{2} + \frac{\ell}{2}(\delta_i + \upsilon_i) - \frac{1}{8}(\delta_i^2 + 6\delta_i\upsilon_i + 3\upsilon_i^2)\right\}. \end{aligned}$$

These expansions remain valid after parameter differentiation near the Gaussian solution for each fixed nonzero $\mathbf{y}$.

E.4.3 Differentiating the fitted parameters

The derivative formula in Theorem A.2, applied to the stationary equations at $\beta = 0$, gives a linear system for the derivatives of the fitted parameters. Use coordinates $(\boldsymbol{m}, \boldsymbol{S}, \eta)$ with $\eta = \log\tau$, and write $\eta^{(\beta)} = \log\widehat{\tau}^{(\beta)}$ along the solution path. At the Gaussian solution, the Hessian $\mathcal{H}$ of J_0 is block diagonal, with the following blocks for any symmetric perturbation $\delta\boldsymbol{S} \in \mathbb{R}^{n\times n}$:

$$\mathcal{H}_{\boldsymbol{m}\boldsymbol{m}} = \widehat{\tau}^{(0)}(\boldsymbol{S}^{(0)})^{-1}, \qquad \mathcal{H}_{\eta\eta} = n/2,$$

$$\text{and} \quad \mathcal{H}_{\boldsymbol{S}\boldsymbol{S}}[\delta\boldsymbol{S}, \delta\boldsymbol{S}] = \frac{1}{2}\operatorname{Tr}((\boldsymbol{S}^{(0)})^{-1}\delta\boldsymbol{S}(\boldsymbol{S}^{(0)})^{-1}\delta\boldsymbol{S}).$$

Here $\mathcal{H}_{\boldsymbol{S}\boldsymbol{S}}[\delta\boldsymbol{S}, \delta\boldsymbol{S}]$ is the quadratic form of the $\boldsymbol{S}$ block in direction $\delta\boldsymbol{S}$. The Hessian is nondegenerate, so this system determines the fitted parameter derivatives from the gradient of J_1. To express these derivatives, introduce the smoother matrix $\boldsymbol{H} = \boldsymbol{K}\boldsymbol{K}_y^{-1}$ and residual matrix $\boldsymbol{P} = \sigma^2\boldsymbol{K}_y^{-1} = \boldsymbol{I}_n - \boldsymbol{H}$. Write $h_i = [\boldsymbol{H}]_{ii}$ and $p_i = [\boldsymbol{P}]_{ii}$ for their diagonal entries. Write $\mathbf{e}^{(\beta)} = \mathbf{y} - \boldsymbol{m}^{(\beta)}$ and $\mathrm{e}_i^{(\beta)} = [\mathbf{e}^{(\beta)}]_i$ for the fitted residual vector and its entries. Equation 20 gives $\boldsymbol{S}^{(0)} = \sigma^2\boldsymbol{H}$, $\mathbf{e}^{(0)} = \mathbf{y} - \boldsymbol{m}^{(0)} = \boldsymbol{P}\mathbf{y}$, and

$$Q_{\mathbf{y}} = \|\mathbf{e}^{(0)}\|_2^2/\sigma^2 + (\boldsymbol{m}^{(0)})^\top\boldsymbol{K}^{-1}\boldsymbol{m}^{(0)}.$$

Define $r_i^2 = \widehat{\tau}^{(0)}(\mathrm{e}_i^{(0)})^2/\sigma^2$, $\ell = \frac{1}{2}\log\{\widehat{\tau}^{(0)}/(2\pi\sigma^2)\}$, and $\chi_i = \ell - 3(h_i + r_i^2)/2$. Solving the differentiated stationarity equations using ∇J_1 gives

$$\partial_\beta\boldsymbol{m}^{(\beta)}\Big|_{0+} = \sigma^{-2}\boldsymbol{S}^{(0)}\operatorname{diag}\left\{\ell - \frac{1}{2}(r_i^2 + 3h_i)\right\}\mathbf{e}^{(0)}, \tag{36}$$

$$\partial_\beta\boldsymbol{S}^{(\beta)}\Big|_{0+} = -\sigma^{-2}\boldsymbol{S}^{(0)}\operatorname{diag}(\chi_i)\boldsymbol{S}^{(0)}, \tag{37}$$

$$\partial_\beta\eta^{(\beta)}\Big|_{0+} = -1 - \frac{\nu_{\mathrm{sm}} + \sum_i r_i^2}{2n} + \frac{1}{2n}\sum_i r_i^2(3h_i + r_i^2) + \frac{\ell}{n}\left(n - \sum_i r_i^2\right). \tag{38}$$

This calculation includes the prior penalty $\tau\boldsymbol{m}^\top\boldsymbol{K}^{-1}\boldsymbol{m}/2$ on the variational mean.

E.4.4 Averaging over the training direction

To average the fitted parameter derivatives over training samples, apply Theorem A.4 to $\sqrt{\tau_0}\boldsymbol{K}_y^{-1/2}\mathbf{y}$:

$$\mathbf{y} = \sqrt{\mathrm{w}/\tau_0}\,\boldsymbol{K}_y^{1/2}\mathbf{v},$$

$$\text{where} \quad \mathrm{w} \sim \chi^2(\mathrm{w}; n), \quad \mathbf{v} \sim \mathrm{Unif}(\mathbb{S}^{n-1}), \quad \text{and} \quad \mathrm{w} \perp \mathbf{v}.$$

Here $\mathbb{S}^{n-1} = \{\boldsymbol{v} \in \mathbb{R}^n : \|\boldsymbol{v}\|_2 = 1\}$ is the unit sphere, and $\mathrm{Unif}(\mathbb{S}^{n-1})$ is its uniform distribution. The quantities r_i^2 depend only on $\mathbf{v}$. The second-moment identity in Theorem A.4 gives $\mathbb{E}_{\mathbf{v}}[r_i^2] = p_i$, with $\sum_i p_i = n - \nu_{\mathrm{sm}}$. The coefficient of ℓ in the directional average of equation 38 is therefore ν_{sm}/n, so

$$\mathbb{E}_{\mathbf{v}}[\,\partial_\beta \eta^{(\beta)}\Big|_{0+} \mid \mathrm{w}] = \frac{\nu_{\mathrm{sm}}}{n}\ell + b_\eta. \tag{39}$$

Here b_η is finite and independent of w: the remaining terms are polynomials of the direction on the compact unit sphere.

E.4.5 Change in the predictive scale matrix

Define the matrix that determines the contribution from changes in $\boldsymbol{V}_*^{(\beta)}$:

$$\boldsymbol{\Lambda}_* = \sigma^2 \boldsymbol{V}_*^{-1/2} \boldsymbol{K}_{X*}^\top \boldsymbol{K}_y^{-2} \boldsymbol{K}_{X*} \boldsymbol{V}_*^{-1/2}, \quad \text{with} \quad \mathrm{Tr}(\boldsymbol{\Lambda}_*) = h(\boldsymbol{X}_*).$$

Because $\boldsymbol{K}_{X*}^\top \boldsymbol{K}_y^{-1}$ maps training responses to the Gaussian predictive mean, the identity

$$h(\boldsymbol{X}_*) = \sigma^2 \|\boldsymbol{V}_*^{-1/2} \boldsymbol{K}_{X*}^\top \boldsymbol{K}_y^{-1}\|_{\mathrm{F}}^2$$

expresses the test input contribution as the squared prediction weights after normalization by the conditional covariance, where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. Using $\boldsymbol{S}^{(0)} \boldsymbol{A}_*^\top = \sigma^2 \boldsymbol{K}_y^{-1} \boldsymbol{K}_{X*}$ in equation 37 gives

$$\mathbb{E}_{\mathbf{v}}\left[\boldsymbol{V}_*^{-1/2}\, \partial_\beta \boldsymbol{V}_*^{(\beta)}\Big|_{0+} \boldsymbol{V}_*^{-1/2} \,\middle|\, \mathrm{w}\right] = -\boldsymbol{\Lambda}_* \left(\ell - \frac{3}{2}\right).$$

For $\boldsymbol{\Sigma}_*^{(\beta)} = \boldsymbol{V}_*^{(\beta)}/\widehat{\tau}^{(\beta)}$, define $\boldsymbol{M}$ by normalizing its derivative using the scale matrix at $\beta = 0$:

$$\begin{aligned} \boldsymbol{M} &= (\boldsymbol{\Sigma}_*^{(0)})^{-1/2}\, \partial_\beta \boldsymbol{\Sigma}_*^{(\beta)}\Big|_{0+} (\boldsymbol{\Sigma}_*^{(0)})^{-1/2} \\ &= \boldsymbol{V}_*^{-1/2}\, \partial_\beta \boldsymbol{V}_*^{(\beta)}\Big|_{0+} \boldsymbol{V}_*^{-1/2} - \partial_\beta \eta^{(\beta)}\Big|_{0+} \boldsymbol{I}_p. \end{aligned}$$

Combining this identity with equation 39 gives

$$\mathbb{E}_{\mathbf{v}}[\boldsymbol{M} \mid \mathrm{w}] = \frac{1}{2}\left(\frac{\nu_{\mathrm{sm}}}{n}\boldsymbol{I}_p + \boldsymbol{\Lambda}_*\right) \log \mathrm{w} + \boldsymbol{M}_c. \tag{40}$$

The matrix $\boldsymbol{M}_c$ is finite and independent of w, by the bounded directional polynomials used in equation 39.

E.4.6 Derivative of the expected log predictive density

The contribution of $\boldsymbol{M}$ to the derivative of the Student-t log density is expressed using the matrix $\boldsymbol{U}_{n,p}$ below. For $\mathbf{t} = \sqrt{n/\mathrm{w}}\,\mathbf{z}$, define

$$\boldsymbol{U}_{n,p}(\mathbf{t}) = \frac{1}{2}\left\{\frac{n+p}{n + \|\mathbf{t}\|_2^2}\mathbf{t}\mathbf{t}^\top - \boldsymbol{I}_p\right\}.$$

The term due to changes in the predictive mean is odd in the standardized test error $\mathbf{z}$. Conditional on the training responses, $\mathbf{z}$ is Gaussian with mean zero. Both $\|(\boldsymbol{\Sigma}_*^{(0)})^{-1/2}\, \partial_\beta \boldsymbol{\mu}_*^{(\beta)}\Big|_{0+}\|_2$ and the operator norm $\|\boldsymbol{M}\|_2$ are bounded by a fixed multiple of $1 + |\log \mathrm{w}|$. These bounds ensure integrability, so Theorem A.5, applied conditionally on the training responses, gives zero conditional expectation for this term. Theorem A.3 and the integrable bound in equation 35 permit differentiation under the expectation:

$$\frac{\partial}{\partial \beta}\mathcal{R}_p(q_\beta^{(p)})\Big|_{0+} = -\mathbb{E}[\mathrm{Tr}\{\boldsymbol{U}_{n,p}(\mathbf{t})\boldsymbol{M}\}]. \tag{41}$$

Averaging over the Gaussian direction reduces the matrix expectation to a beta integral. Let $\mathrm{u} = \|\mathbf{z}\|_2^2$ and $\mathrm{b} = \mathrm{u}/(\mathrm{w} + \mathrm{u})$. By Theorem A.7, $\mathrm{b} \sim \mathrm{Beta}(p/2, n/2)$ is independent of $\mathrm{w} + \mathrm{u}$, and the conditional second moment in Theorem A.4, applied to $\mathbf{z}$, gives

$$\mathbb{E}[\boldsymbol{U}_{n,p}(\mathbf{t}) \mid \mathrm{w}, \mathrm{u}] = \frac{1}{2}\left(\frac{n+p}{p}\mathrm{b} - 1\right)\boldsymbol{I}_p.$$

To compute the required logarithmic moment, temporarily write the beta density as $f_{\alpha,\gamma}(x) = x^{\alpha-1}(1-x)^{\gamma-1}/B(\alpha,\gamma)$ and let $\mathbb{E}_{\alpha,\gamma}$ denote expectation under this density. Differentiating the normalizing integral gives $\partial_\gamma \log B(\alpha,\gamma) = \mathbb{E}_{\alpha,\gamma}[\log(1-\mathrm{b})]$, so

$$\partial_\gamma f_{\alpha,\gamma}(x) = f_{\alpha,\gamma}(x)\big\{\log(1-x) - \mathbb{E}_{\alpha,\gamma}[\log(1-\mathrm{b})]\big\}.$$

Theorem A.3 therefore turns the derivative of the beta mean into a covariance:

$$\begin{aligned}\partial_\gamma \frac{\alpha}{\alpha+\gamma} &= \int_0^1 x\,\partial_\gamma f_{\alpha,\gamma}(x)\,dx \\ &= \mathbb{E}_{\alpha,\gamma}[\mathrm{b}\log(1-\mathrm{b})] - \mathbb{E}_{\alpha,\gamma}[\mathrm{b}]\,\mathbb{E}_{\alpha,\gamma}[\log(1-\mathrm{b})] = -\frac{\alpha}{(\alpha+\gamma)^2}.\end{aligned}$$

At $(\alpha,\gamma) = (p/2, n/2)$, the mean is $p/(n+p)$ and the covariance is $-2p/(n+p)^2$. For this application, integrate against the fixed uniform probability measure on $(0,1)$ and restrict the shape parameter to a compact neighborhood of $n/2$ inside $(0,\infty)$. The beta-density derivatives are then dominated by integrable powers of b and $1-\mathrm{b}$ times $1+|\log(1-\mathrm{b})|^2$. Using $\log \mathrm{w} = \log(1-\mathrm{b}) + \log(\mathrm{w}+\mathrm{u})$, we obtain

$$\mathbb{E}[\boldsymbol{U}_{n,p}(\mathbf{t})] = \mathbf{0}_{p\times p}, \quad \text{and} \quad \mathbb{E}[\boldsymbol{U}_{n,p}(\mathbf{t})\log \mathrm{w}] = -\frac{1}{n+p}\boldsymbol{I}_p.$$

Theorem A.4 and the independence of the conditional Gaussian error imply that the training direction is independent of $(\mathrm{w}, \mathbf{z})$. Inserting equation 40 therefore cancels $\boldsymbol{M}_c$ and gives

$$\frac{\partial}{\partial\beta}\mathcal{R}_p(q_\beta^{(p)})\Big|_{0+} = \frac{\mathrm{Tr}\{(\nu_{\mathrm{sm}}/n)\boldsymbol{I}_p + \boldsymbol{\Lambda}_*\}}{2(n+p)} = c_{1,p}. \tag{42}$$

E.5 Remainder bound and gain retention

The upper bound $H_{2,p}(\mathbf{y})$ in equation 35 supplies the remainder constant

$$\overline{C}_{n,p} := \frac{1}{2}\mathbb{E}[H_{2,p}(\mathbf{y})] \le C(n+p)d_0^2\mathbb{E}\big[L_{\mathbf{y}}^2(1+e^{b_0 L_{\mathbf{y}}})\big] < \infty. \tag{43}$$

For a fixed training setup, this bound is proportional to $n+p$; its remaining dependence is through d_0 and the distribution of $L_{\mathbf{y}}$, which is determined by n, s, τ_0. For $L_p(\beta) = -\log q_\beta^{(p)}(\mathbf{y}_* \mid \mathbf{y})$, Theorem A.6 gives

$$L_p(\beta) - L_p(0) - \beta\,\partial_b L_p(b)|_{b=0^+} = \int_0^\beta (\beta-b)\partial_b^2 L_p(b)\,db,$$

where $\partial_b = \partial/\partial b$. The bound in equation 35 makes the absolute remainder at most $\beta^2 H_{2,p}(\mathbf{y})/2$ for each nonzero training vector. Taking expectations of this bound and using equation 42 yields

$$\Big|\mathcal{R}_p(q_\beta^{(p)}) - \mathcal{R}_p(q_0^{(p)}) - c_{1,p}\beta\Big| \le \overline{C}_{n,p}\beta^2 \quad \text{for } 0 \le \beta \le b_0. \tag{44}$$

Subtracting this cost from the exact gain in Lemma 1 proves equation 10. The condition $c_{1,p}\beta + \overline{C}_{n,p}\beta^2 \le \rho\Delta_{n,p}$ then gives the stated fraction of retained gain. Because $\Delta_{n,p} > 0$ and the cost bound tends to zero with β, this condition holds on a nonzero interval for each $\rho \in (0,1)$. This completes the proof of Theorem 1. □

F Experimental details and supplementary results

The details follow the two experiments in the main text: data generated from the assumed GP model, and prediction after adding outliers to the training responses.

F.1 Numerical implementation

Synthetic data & Model setup. Synthetic training and test responses were generated from a zero mean GP with an RBF kernel $k(\boldsymbol{x},\boldsymbol{x}') = \exp\{-\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2/(2\ell_k^2)\}$. The shared precision was $\tau_0 = 1$, the SD of the observation noise was $\sigma = 0.2$, and the lengthscale was $\ell_k = 0.15$ on $[0,1]$. We evaluated training sample sizes $n \in \{4,6,8,12,16,32,64\}$ and prediction dimensions $p \in \{1,2,4,8,16,32,64,128\}$. Training inputs were equally spaced cell midpoints, and test inputs

were the first p points of a common shifted Sobol sequence (Sobol', 1967). Each design condition used 500 independent data replicates. Both experiments optimized the DPD objective in equation 7, with the true kernel parameters and relative noise variance held fixed.

Optimization protocol. For experiments on uncontaminated data, DPD optimization was initialized at the exact Gaussian solution ($\beta = 0$) and continued sequentially across increasing β values. Each step used L-BFGS-B (Byrd et al., 1995). We periodically recentered and rescaled the optimization coordinates around the current fit to maintain good numerical conditioning. A diagonal kernel regularization of 10^{-8} was applied consistently. For experiments on contaminated data, we initialized optimization from the Gaussian fit, the fit at the preceding β, and the prior mean and covariance with $\tau = 1$. We selected the accepted solution with the lowest objective value.

Numerical stability & Reproducibility. All accepted fits used to compare predictive scores by Monte Carlo met the following numerical conditions: positive precision, positive definite covariance, and a preconditioned full gradient whose maximum norm was at most 2×10^{-6}. Fits obtained from contaminated data also met this tolerance for the relative precision residual. Estimated errors in the conditional quadrature remained below 2×10^{-7}, and our numerical integration agreed with the exact Studentization gain within 10^{-11} across the full (n, p) design. All Hessians evaluated at the fitted solutions in the Monte Carlo study of the second-order correction were positive definite. Note that these checks ensure numerical stationarity and accuracy, but do not formally identify the solutions with the local solutions constructed in Theorem 1. Complete random seeds, software environment details, and execution logs are provided in the code supplement's `README.md`.

F.2 Finite-sample gain under the assumed GP model

F.2.1 Gain estimation

We estimated the predictive cost of DPD fitting by Monte Carlo, using a control variate to reduce variance. Subtracting this estimate from the exact Studentization gain $\Delta_{n,p}$ gave our estimate of $G_p(\beta)$. We evaluated a fixed grid $\beta \in \{0, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.025, 0.05, 0.1, 0.2\}$ before data generation.

For a training vector $\mathbf{y}$, the conditional predictive cost of DPD fitting and its right derivative at $\beta = 0$ are:

$$C_\beta(\mathbf{y}) = \mathbb{E}\left[\log \frac{q_0^{(p)}(\mathbf{y}_* \mid \mathbf{y})}{q_\beta^{(p)}(\mathbf{y}_* \mid \mathbf{y})} \,\middle|\, \mathbf{y}\right], \quad \text{and} \quad \dot{C}_0(\mathbf{y}) = \left.\frac{\partial C_\beta(\mathbf{y})}{\partial \beta}\right|_{0+}.$$

The integral over test responses was reduced to one dimensional quadrature using the conditional Gaussian distribution. For this computation, let $\boldsymbol{Z}$ be the conditional test error standardized by the fitted predictive scale, with covariance eigenvalues λ_j and mean coordinates u_j. We evaluated:

$$\mathbb{E}\left[\log\left(1 + \frac{\|\boldsymbol{Z}\|_2^2}{n}\right)\right] = \int_0^\infty \frac{e^{-s}}{s}\{1 - L(s/n)\}\, ds,$$
$$\text{where} \quad L(v) = \prod_{j=1}^{p}(1 + 2v\lambda_j)^{-1/2} \exp\left\{-\sum_{j=1}^{p}\frac{v u_j^2}{1 + 2v\lambda_j}\right\}.$$

To reduce sampling variability, we used a control variate based on an independently estimated first-order coefficient. We used an independent sample of 65,536 training vectors per design to estimate $\mathbb{E}[\dot{C}_0(\mathbf{y})]$ by $\widehat{c}_{\text{ind}}$. The final control variate estimator across $M = 500$ replicates was:

$$\widehat{C}_\beta = \frac{1}{M}\sum_{m=1}^{M}\left\{C_\beta(\mathbf{y}^{(m)}) - \beta \dot{C}_0(\mathbf{y}^{(m)})\right\} + \beta \widehat{c}_{\text{ind}}, \tag{45}$$
$$\text{and} \quad \widehat{G}_p(\beta) = \Delta_{n,p} - \widehat{C}_\beta.$$

F.2.2 First-order cost and gains at the tested β values

Figure 2 in the main text compares the estimated gain with the first-order approximation, and Figure 5 shows the full design at selected β values.

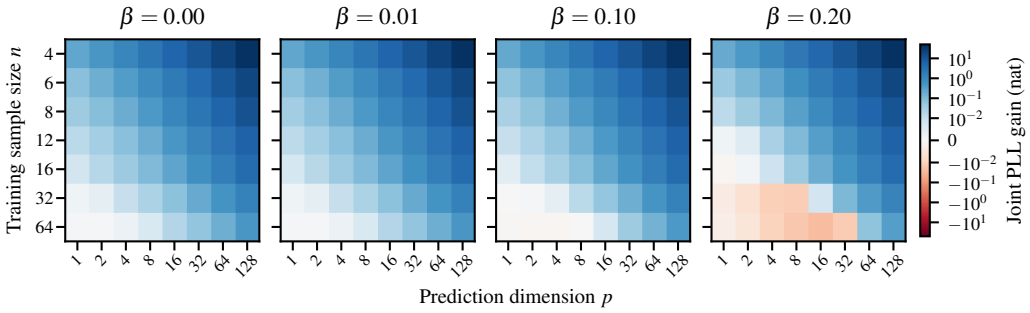


Figure 5: Gain over ordinary Gaussian prediction across the full (n, p) design. All estimated gains were positive at $\beta = 0.01$; negative gains first appear at small p as β increases. Table 2 reports representative estimates and SEs.

We independently checked the first-order coefficient $c_{1,p}$ against the sampled derivative estimate. The theoretical coefficient lay within the pointwise 95% interval for the initial derivative estimate in 48 of the 56 (n, p) conditions. We rechecked the 8 conditions with larger standardized differences using 262,144 new training vectors; all had absolute standardized differences of at most 0.56. The reported gain estimates retain the original derivative sample. Table 2 gives representative gain estimates.

Table 2: Joint PLL gain over the ordinary Gaussian predictor on uncontaminated synthetic data at $n = 64$. The table reports the exact Studentization gain $\Delta_{64,p}$ and the estimated gain after DPD fitting (Monte Carlo SE), both in 10^{-3} nats.

p	β	$\Delta_{64,p}$	$\widehat{G}_p(\beta)$ (SE)
1	0.01	0.4436	0.4180 (0.0014)
	0.10		−0.5110 (0.0929)
	0.20		−2.8650 (0.4072)
16	0.01	19.8003	19.4643 (0.0149)
	0.10		7.3853 (0.5806)
	0.20		−23.3946 (2.3644)
128	0.01	506.5303	505.5205 (0.0468)
	0.10		461.9763 (1.7999)
	0.20		353.6874 (6.9447)

F.2.3 Estimating the second-order correction and gain retention

We estimated the second-order correction on a separate sample to assess how the local expansion describes gain retention at finite β. For each training sample size, we generated 128 new training vectors. For each fit, we drew 128 vectors of test responses from the conditional distribution. Starting from the Gaussian solution at $\beta = 0$, we fitted the model at successive β values spaced by 0.005 up to $\beta = 0.205$. First and second derivatives were evaluated by finite differences.

Let D^2_{β,h_i} denote the finite difference operator for the second derivative with respect to β, using the fitted models for replicate i. We computed $\widehat{C}_p$ as the largest absolute value of the estimated second derivative over the tested grid. The resulting empirical retention curve and its 50% cutoff were:

$$\widehat{L}_p(\beta) = \Delta_{n,p} - c_{1,p}\beta - \widehat{C}_p\beta^2,$$

$$\text{and} \quad \widehat{\beta}_{50,p} = \min\left\{0.2, \frac{\Delta_{n,p}}{c_{1,p} + \sqrt{c_{1,p}^2 + 2\widehat{C}_p\Delta_{n,p}}}\right\}.$$

We validated this cutoff using gain estimates based on 500 independent replicates from the experiment on uncontaminated data (Section 6.2). Of the 504 conditions with $\beta > 0$, 446 satisfied $\beta \le \widehat{\beta}_{50,p}$. These empirical cutoffs are diagnostics for the tested grid, not certified instances of the interval in Theorem 1 that is independent of the training responses. All independently estimated retention

fractions for these conditions exceeded 50%. The smallest 95% lower confidence bound across these conditions was 62.0%.

Table 3: Empirical 50% retention cutoffs and independent gain estimates at $n = 64$. Absolute gains are reported for the full prediction vector and per response, in 10^{-3} nats. Retention is the gain divided by the exact Studentization gain; the 95% intervals describe Monte Carlo uncertainty.

			Absolute PLL gain (10^{-3} nat)		Retention (%)	
p	$\widehat{\beta}_{50,p}$	Tested β	Joint $\widehat{G}_p$	Per response $\widehat{G}_p/p$	Estimate	95% interval
1	0.0231	0.010	0.4180	0.4180	94.2	[93.6, 94.8]
16	0.0686	0.050	16.0720	1.0045	81.2	[79.6, 82.7]
32	0.1017	0.100	39.2942	1.2279	65.1	[62.0, 68.2]
128	0.2000	0.200	353.6874	2.7632	69.8	[67.1, 72.5]

F.3 Prediction under contamination

F.3.1 Response contamination and score comparisons

Following the uniform scheme of Altamirano et al. (2024, Appendix B.3.2), we added signed shifts to a target fraction $r = 0.1$ of training responses. Specifically, we selected exactly $\max\{1, \lfloor nr + 0.5 \rfloor\}$ points uniformly at random. Because at least one response was contaminated, the actual contamination fraction exceeded 10% at small training sample sizes (e.g., 25% at $n = 4$). Shift magnitudes were drawn uniformly from $(3s_y, 9s_y)$, where s_y is the SD of the uncontaminated training responses. Test responses were uncontaminated.

After DPD fitting, we held the fitted parameters fixed and evaluated the additional gain from Studentization:

$$\mathbb{E}_{\text{cont}}\left[\log \frac{q_{\beta}^{(p)}(\mathbf{y}_* \mid \mathbf{y})}{q_{\text{G}}^{(p)}(\mathbf{y}_* \mid \mathbf{y}; \beta)}\right], \tag{46}$$

where $q_{\text{G}}^{(p)}$ represents the DPD Gaussian predictor. The total gain over ordinary Gaussian prediction was computed using the ordinary fit ($\beta = 0$) on the same contaminated training vector.

F.3.2 Retention and robustness at the same β

We assessed retention and robust fitting simultaneously using $n \in \{8, 16, 64\}$, $p \in \{1, 16, 128\}$, and $\beta \in \{0.01, 0.05, 0.1\}$, with 500 replicates per design. We applied uniform response contamination to the exact clean training draws from the experiment on uncontaminated data (Section 6.2). Test scores were evaluated using the quadrature method (Appendix F.2.1). Figure 3 in the main text summarizes these results.

F.3.3 Training sample size and prediction dimension

We extended the paired Studentization comparison at $\beta = 0.1$ to all 56 combinations of n and p.

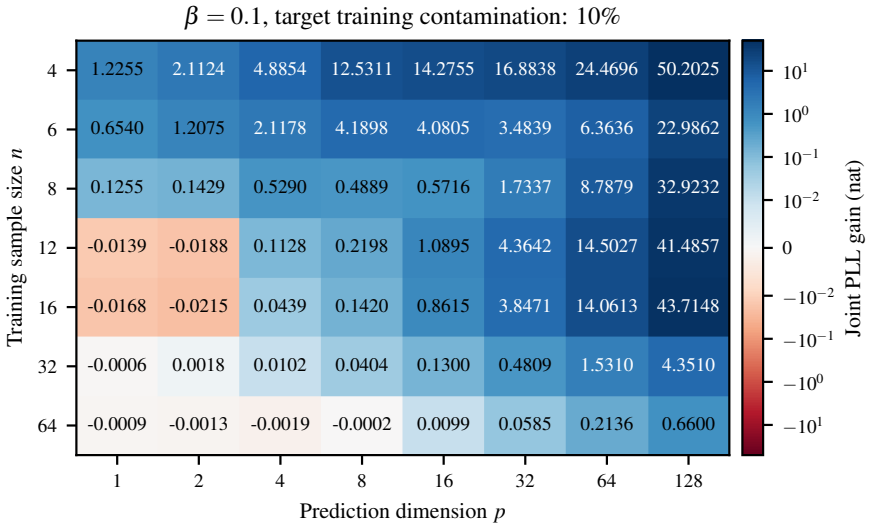


Figure 6: Additional gain from Studentization after DPD fitting at $\beta = 0.1$. Outliers were added to a target fraction $r = 0.1$ of the training responses. Mean gains were positive in 47 of 56 (n, p) combinations, with negative means confined to $p \leq 8$ and $n \geq 12$.

Table 4: Joint PLL gain of the DPD Studentized predictor over the DPD Gaussian predictor from the same fit. Here, $\beta = 0.1$, and outliers were added to a target fraction $r = 0.1$ of the training responses. Each cell shows the mean, with its SE below, across 500 independent replicates.

$n \backslash p$	1	2	4	8	16	32	64	128
4	1.2255 (0.7468)	2.1124 (1.2592)	4.8854 (2.6499)	12.5311 (3.6213)	14.2755 (5.2979)	16.8838 (9.1976)	24.4696 (16.0128)	50.2025 (29.9271)
6	0.6540 (0.0847)	1.2075 (0.1000)	2.1178 (0.1875)	4.1898 (0.3305)	4.0805 (0.4827)	3.4839 (0.7551)	6.3636 (1.3234)	22.9862 (2.5623)
8	0.1255 (0.0257)	0.1429 (0.0251)	0.5290 (0.0684)	0.4889 (0.0710)	0.5716 (0.0864)	1.7337 (0.1006)	8.7879 (0.3243)	32.9232 (0.9826)
12	−0.0139 (0.0028)	−0.0188 (0.0054)	0.1128 (0.0289)	0.2198 (0.0354)	1.0895 (0.0855)	4.3642 (0.2111)	14.5027 (0.5154)	41.4857 (1.3094)
16	−0.0168 (0.0008)	−0.0215 (0.0015)	0.0439 (0.0189)	0.1420 (0.0174)	0.8615 (0.0392)	3.8471 (0.1236)	14.0613 (0.4053)	43.7148 (1.2112)
32	−0.0006 (0.0010)	0.0018 (0.0017)	0.0102 (0.0034)	0.0404 (0.0069)	0.1300 (0.0164)	0.4809 (0.0505)	1.5310 (0.1468)	4.3510 (0.4233)
64	−0.0009 (0.0003)	−0.0013 (0.0005)	−0.0019 (0.0009)	−0.0002 (0.0014)	0.0099 (0.0031)	0.0585 (0.0094)	0.2136 (0.0296)	0.6600 (0.0918)

F.3.4 Comparison of four predictors at fixed β

We isolated the effects of robust fitting and Studentization at $\beta = 0.1$ by evaluating mean pointwise PLL over 128 clean test responses in 500 replicates. The inputs were disjoint regular grids in $[0, 1]$.

Table 5: Mean pointwise PLL with uniform contamination of the training responses, with $\beta = 0.1$ for DPD fitting. Scores are means (SEs) in nats per response across 500 replicates. The last column gives the paired DPD Studentized minus DPD Gaussian difference in 10^{-3} nats.

n	Ordinary Gaussian	Ordinary Studentized	DPD Gaussian	DPD Studentized	Paired gain
4	−3.611 (0.073)	−2.101 (0.022)	−3.191 (0.068)	−2.065 (0.021)	1126.375 (60.707)
6	−2.958 (0.044)	−1.874 (0.022)	−2.504 (0.038)	−1.780 (0.021)	723.810 (29.873)
8	−1.932 (0.028)	−1.586 (0.020)	−1.550 (0.028)	−1.362 (0.023)	188.322 (12.075)
12	−1.425 (0.019)	−1.363 (0.019)	−0.879 (0.023)	−0.855 (0.022)	23.814 (4.841)
16	−1.550 (0.019)	−1.559 (0.019)	−1.039 (0.027)	−1.037 (0.027)	1.977 (4.251)
32	−1.388 (0.017)	−1.397 (0.017)	−0.020 (0.010)	−0.019 (0.010)	0.539 (0.648)
64	−1.343 (0.016)	−1.347 (0.016)	0.093 (0.005)	0.093 (0.004)	−0.032 (0.193)

F.3.5 Sensitivity to contamination

Finally, we varied the fraction and pattern of contamination. At $n = 64$, increasing the uniform contamination fraction from $r = 0.1$ to $r = 0.2$ reduced the scores of all predictors, and the additional gain from Studentization became slightly negative (-0.0044 nat per response).

At $n = 8$, we further compared uniform, focused, and asymmetric contamination mechanisms (Altamirano et al., 2024). Asymmetric contamination exclusively applied negative shifts. Focused contamination relocated the selected input to the larger of the two central training inputs after sorting and then replaced its response to mimic a cluster of anomalies (Figure 7).

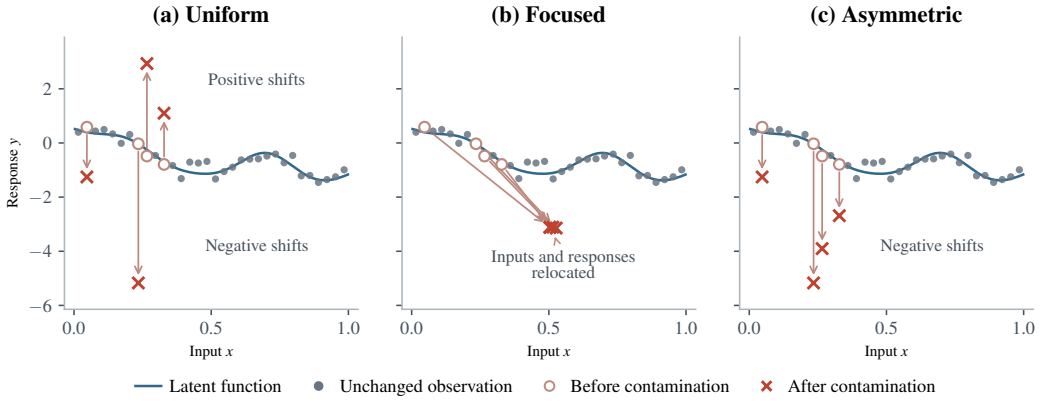


Figure 7: Uniform contamination shifts responses in both directions; asymmetric contamination shifts them downward; focused contamination relocates inputs and responses together.

Table 6: Mean pointwise PLL across contamination patterns at $n = 8$, with one contaminated observation and fixed $\beta = 0.1$. Scores are means (SEs) across 500 replicates, in nats per response. The last column gives the paired DPD Studentized minus DPD Gaussian difference in 10^{-3} nats.

Pattern	Ordinary Gaussian	Ordinary Studentized	DPD Gaussian	DPD Studentized	Paired gain
Uniform	−1.932 (0.028)	−1.586 (0.020)	−1.550 (0.028)	−1.362 (0.023)	188.322 (12.075)
Focused	−1.022 (0.016)	−1.033 (0.015)	−0.992 (0.016)	−1.012 (0.015)	−20.505 (4.637)
Asymmetric	−1.941 (0.026)	−1.599 (0.019)	−1.533 (0.029)	−1.347 (0.024)	185.699 (13.385)

Studentization increased the gains from DPD fitting under both uniform and asymmetric contamination, but under severe focused contamination, it slightly reduced PLL (Table 6).

G Additional results for existing robust GPs

G.1 Data and contamination

The dataset sizes and corresponding training and test sizes are listed in Table 7. The real datasets from the UCI Machine Learning Repository (Asuncion & Newman, 2007) were Boston Housing (Harrison & Rubinfeld, 1978), Energy Efficiency (Tsanas & Xifara, 2012), and Yacht Hydrodynamics (Gerritsma et al., 1981). The synthetic dataset follows the released RCGP generator: 300 equally

Table 7: Datasets for the comparison before and after Studentization.

Dataset	Total observations	Training n	Test p
Synthetic	300	270	30
Boston	506	455	51
Energy	768	691	77
Yacht	308	277	31

spaced inputs on $[0, 20]$, an RBF kernel with unit variance and lengthscale, and Gaussian observation noise with SD 0.1. We used the same generated dataset across splits, so the repetitions measure sensitivity to the split rather than variation across independent GP draws. The Energy target was heating load, with the eight original predictors.

For each contaminated condition, we modified $\lfloor 0.1n \rfloor$ randomly selected training observations using the released RCGP mechanisms. Uniform contamination adds shifts of either sign, with magnitudes between three and nine times the SD of the training responses. Asymmetric contamination uses the same magnitude range with a common randomly chosen sign within each contaminated sample. Focused contamination relocates selected inputs to a small cluster near the medians of the input coordinates. Their responses are replaced with values below a central response level by approximately three times the SD of the responses. Test observations were left unmodified.

In the main evaluation, responses were standardized using the possibly contaminated training data, and the real data inputs were standardized using training statistics. The synthetic inputs remained in their original units. All three methods received the same transformed arrays, without access to the contamination labels or the original clean training responses. Predictive scores and interval widths were converted back to the original response units.

G.2 Model fitting and prediction

RCGP used an isotropic RBF kernel, a fixed robustness threshold given by the 95th percentile of absolute training responses, and its analytic leave-one-out (LOO) objective for hyperparameter fitting. Relevance Pursuit used the official BoTorch implementation (Balandat et al., 2020) with backward pursuit, maximum a posteriori (MAP) support selection, and a Matérn-5/2 kernel with automatic relevance determination (ARD). The prior on the number of observations assigned additional noise had mean $0.2n$. Each L-BFGS-B call had a budget of 1,024 iterations. The mixture GP used the full Gaussian variational factor of the Gaussian/uniform model and a $\mathrm{Beta}(10, 10)$ prior on the mixture proportion. We fitted the model for 499 iterations, with clipping applied to the momentum terms in the kernel parameter updates. Its kernel and noise parameters were updated along with the variational factors. For the clean test targets, its predictive distribution included Gaussian observation noise for the inlier component.

Studentization was applied after fitting, with one auxiliary precision τ shared across the test vector:

$$\tau \sim \mathrm{Gam}(n/2, n/2),$$
$$\mathbf{Y}_* \mid \tau \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\tau).$$

The Gamma distribution is parameterized by shape and rate, so the parameters above give mean precision one. This mixture defines the predictive correction; it is not assumed to be the precision posterior of the original robust model. Because the predictive mean is unchanged, root mean squared error (RMSE) and mean absolute error (MAE) are identical before and after Studentization.

G.3 Variation across conditions and splits

Figure 8 shows the means, medians, and interquartile ranges for all 16 conditions.

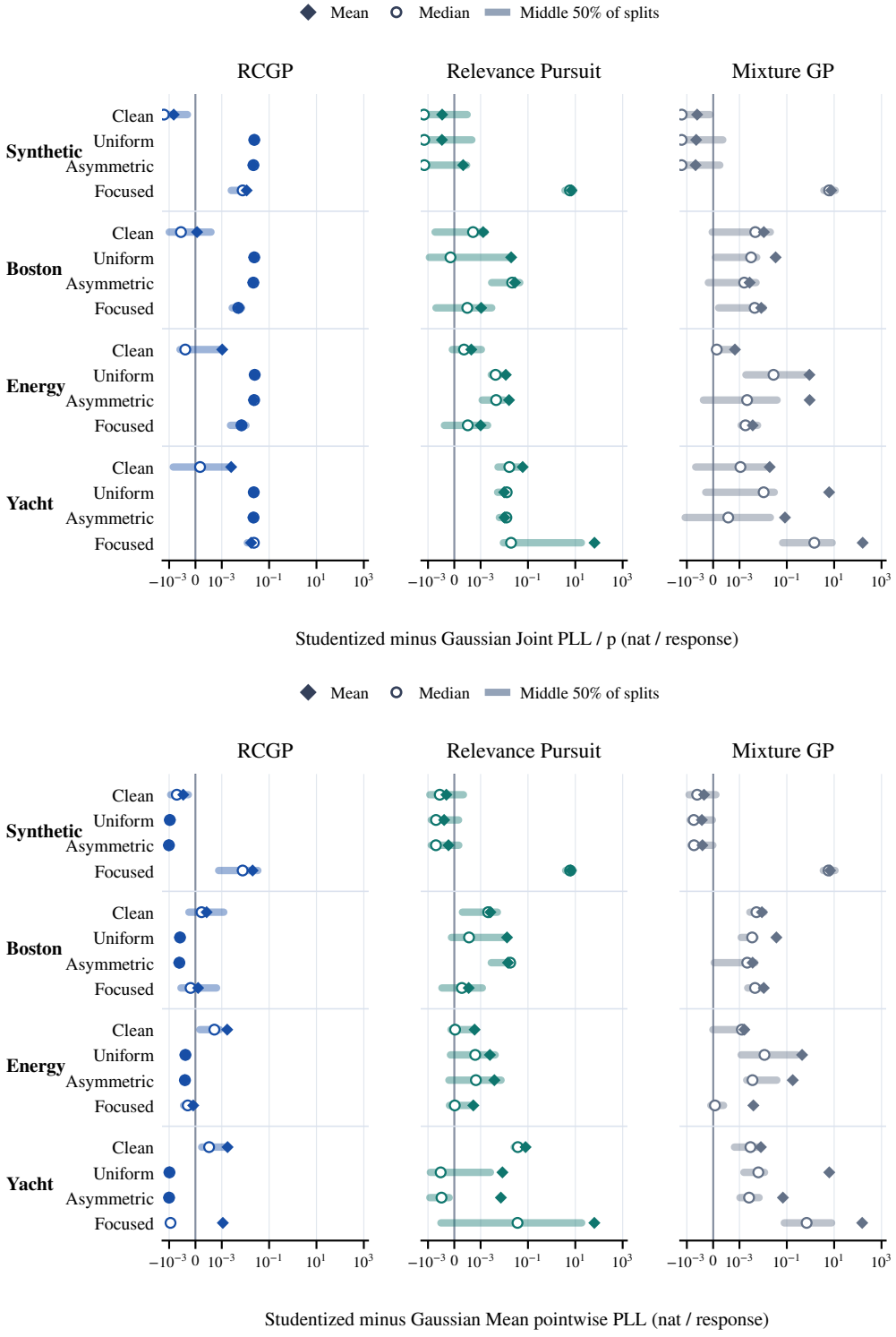


Figure 8: Changes in joint PLL per response (top) and mean pointwise PLL (bottom) across all 16 conditions. Diamonds show means, open circles show medians, and thick segments show the middle 50% of ten paired splits. Positive values favor Studentization; all panels share the horizontal scale used in Figure 4.

Tables 9, 10, and 11 report paired differences separately for each condition, with SEs across splits, medians, and counts of positive differences. The splits share observations, so these SEs describe split variability rather than uncertainty across independent datasets. Improvement counts at the individual split level are given in Table 8. The largest decreases within individual splits were below 0.0013 nat per response for pointwise PLL and 0.0018 for joint PLL per response. Some large mean gains arose from splits with extremely low Gaussian PLL, particularly under focused contamination on Yacht. We therefore report absolute scores and split medians alongside the mean differences.

Table 8: Numbers of splits in which Studentization increased PLL, out of 160 main splits for each model.

Model	Mean pointwise PLL	Joint PLL
RCGP	41/160	134/160
Relevance Pursuit	94/160	108/160
Mixture GP	120/160	105/160

Table 9: RCGP: changes after Studentization across all 16 conditions. Scores are in 10^{-3} nats per response. Each mean, median, and positive count summarizes ten paired splits; SE describes split variability. The $p = 16$ column uses joint PLL divided by 16, averaged over complete, nonoverlapping blocks of 16 test observations.

Mean pointwise PLL

Data	Contamination	Mean (SE)	Median	Positive
Synthetic	clean	−0.464 (0.228)	−0.710	2/10
	uniform	−0.972 (0.008)	−0.974	0/10
	asymmetric	−1.035 (0.013)	−1.022	0/10
	focused	20.096 (8.418)	7.627	9/10
Boston	clean	0.424 (0.264)	0.231	5/10
	uniform	−0.591 (0.005)	−0.587	0/10
	asymmetric	−0.623 (0.008)	−0.611	0/10
	focused	0.100 (0.255)	−0.185	4/10
Energy	clean	1.695 (1.182)	0.720	8/10
	uniform	−0.383 (0.003)	−0.382	0/10
	asymmetric	−0.405 (0.005)	−0.400	0/10
	focused	−0.084 (0.179)	−0.294	2/10
Yacht	clean	1.756 (0.998)	0.519	9/10
	uniform	−0.989 (0.010)	−0.989	0/10
	asymmetric	−1.016 (0.013)	−1.008	0/10
	focused	1.110 (1.386)	−0.948	2/10

Joint PLL / p

Data	Contamination	All test observations			Blocks of 16
		Mean (SE)	Median	Positive	Mean (SE)
Synthetic	clean	−0.826 (0.468)	−1.674	3/10	−0.839 (0.471)
	uniform	23.669 (0.250)	23.654	10/10	12.002 (0.127)
	asymmetric	21.549 (0.467)	21.769	10/10	10.931 (0.281)
	focused	11.253 (4.908)	7.627	8/10	29.031 (14.798)
Boston	clean	0.061 (0.451)	−0.551	4/10	0.387 (0.307)
	uniform	23.228 (0.440)	23.610	10/10	6.724 (0.145)
	asymmetric	21.366 (0.589)	21.599	10/10	6.119 (0.189)
	focused	5.144 (1.079)	4.958	10/10	1.643 (0.323)
Energy	clean	1.046 (0.905)	−0.388	4/10	0.176 (0.178)
	uniform	24.517 (0.206)	24.611	10/10	4.669 (0.048)
	asymmetric	22.364 (0.801)	23.260	10/10	4.194 (0.196)
	focused	6.829 (1.647)	6.806	10/10	1.395 (0.282)
Yacht	clean	2.495 (2.355)	0.175	5/10	0.003 (0.916)
	uniform	22.462 (0.332)	22.478	10/10	10.910 (0.205)
	asymmetric	21.494 (0.508)	21.998	10/10	10.082 (0.395)
	focused	17.323 (3.123)	22.723	10/10	9.609 (1.253)

Table 10: Relevance Pursuit: changes after Studentization across all 16 conditions. Scores are in 10^{-3} nats per response. Each mean, median, and positive count summarizes ten paired splits; SE describes split variability. The $p = 16$ column uses joint PLL divided by 16, averaged over complete, nonoverlapping blocks of 16 test observations.

Mean pointwise PLL

Data	Contamination	Mean (SE)	Median	Positive
Synthetic	clean	−0.301 (0.266)	−0.553	3/10
	uniform	−0.387 (0.258)	−0.696	3/10
	asymmetric	−0.222 (0.387)	−0.699	3/10
	focused	6428.949 (1364.784)	6020.615	9/10
Boston	clean	2.589 (0.844)	2.086	8/10
	uniform	13.057 (9.692)	0.560	7/10
	asymmetric	14.350 (3.564)	17.719	10/10
	focused	0.546 (0.391)	0.287	6/10
Energy	clean	0.771 (0.588)	0.031	6/10
	uniform	2.473 (1.057)	0.794	6/10
	asymmetric	3.809 (1.795)	0.820	6/10
	focused	0.723 (0.589)	0.018	5/10
Yacht	clean	79.155 (47.983)	37.212	10/10
	uniform	8.342 (7.256)	−0.519	3/10
	asymmetric	7.174 (7.396)	−0.487	2/10
	focused	63615.052 (41627.354)	36.203	7/10

Joint PLL / p

Data	Contamination	All test observations			Blocks of 16
		Mean (SE)	Median	Positive	Mean (SE)
Synthetic	clean	−0.462 (0.582)	−1.486	3/10	−0.906 (0.522)
	uniform	−0.470 (0.576)	−1.442	3/10	−0.822 (0.420)
	asymmetric	0.336 (1.290)	−1.430	3/10	−0.609 (0.480)
	focused	6876.928 (1675.794)	5946.407	9/10	9006.299 (3737.961)
Boston	clean	1.280 (0.949)	0.711	6/10	2.754 (1.402)
	uniform	19.505 (12.605)	−0.146	5/10	16.155 (10.187)
	asymmetric	28.172 (9.152)	21.499	8/10	23.308 (7.506)
	focused	1.055 (0.598)	0.494	5/10	0.768 (0.341)
Energy	clean	0.646 (0.393)	0.368	7/10	0.411 (0.175)
	uniform	11.578 (7.358)	4.269	10/10	2.356 (1.485)
	asymmetric	15.836 (8.959)	4.519	8/10	3.259 (1.703)
	focused	1.003 (0.525)	0.512	6/10	0.346 (0.234)
Yacht	clean	59.884 (38.848)	16.339	9/10	113.071 (82.709)
	uniform	9.861 (2.164)	12.678	8/10	11.034 (7.765)
	asymmetric	10.245 (1.923)	12.279	9/10	10.300 (7.951)
	focused	64457.446 (42039.136)	19.301	9/10	72460.284 (72398.583)

Table 11: Mixture GP: changes after Studentization across all 16 conditions. Scores are in 10^{-3} nats per response. Each mean, median, and positive count summarizes ten paired splits; SE describes split variability. The $p = 16$ column uses joint PLL divided by 16, averaged over complete, nonoverlapping blocks of 16 test observations.

Mean pointwise PLL

Data	Contamination	Mean (SE)	Median	Positive
Synthetic	clean	−0.347 (0.249)	−0.609	3/10
	uniform	−0.429 (0.244)	−0.737	3/10
	asymmetric	−0.407 (0.243)	−0.729	3/10
	focused	6715.551 (1681.923)	5783.605	9/10
Boston	clean	8.877 (3.107)	5.208	10/10
	uniform	36.338 (33.373)	3.434	9/10
	asymmetric	3.588 (1.548)	2.125	8/10
	focused	10.646 (4.660)	4.541	9/10
Energy	clean	1.607 (0.685)	1.263	7/10
	uniform	441.839 (331.982)	11.394	8/10
	asymmetric	178.238 (117.064)	3.550	10/10
	focused	3.878 (3.722)	0.068	6/10
Yacht	clean	8.036 (4.116)	2.910	8/10
	uniform	6212.194 (6198.265)	6.272	8/10
	asymmetric	68.460 (65.040)	2.545	9/10
	focused	152837.082 (101212.155)	687.796	10/10

Joint PLL / p

Data	Contamination	All test observations			Blocks of 16
		Mean (SE)	Median	Positive	Mean (SE)
Synthetic	clean	−0.609 (0.583)	−1.653	3/10	−0.793 (0.554)
	uniform	−0.645 (0.535)	−1.663	3/10	−0.683 (0.498)
	asymmetric	−0.668 (0.541)	−1.679	3/10	−0.679 (0.498)
	focused	7500.527 (2043.816)	6116.660	9/10	10056.085 (4159.721)
Boston	clean	10.709 (4.029)	4.651	7/10	11.728 (3.769)
	uniform	33.662 (31.251)	3.096	7/10	41.198 (37.596)
	asymmetric	2.692 (1.085)	1.586	7/10	4.096 (1.672)
	focused	8.330 (3.571)	4.379	8/10	10.932 (4.225)
Energy	clean	0.828 (0.532)	0.137	8/10	0.684 (0.515)
	uniform	901.118 (669.161)	27.442	8/10	573.977 (422.842)
	asymmetric	921.079 (615.360)	2.100	6/10	210.368 (139.034)
	focused	3.630 (1.221)	1.790	9/10	3.937 (3.618)
Yacht	clean	18.935 (13.016)	1.098	7/10	19.842 (17.485)
	uniform	6123.185 (6092.825)	10.510	6/10	6.655 (5.278)
	asymmetric	83.598 (77.978)	0.574	5/10	147.202 (145.400)
	focused	160256.621 (105496.617)	1450.414	9/10	160403.427 (159787.155)

G.4 PREDICTIVE INTERVALS

Tables 12, 13, and 14 report absolute PLL, marginal 95% coverage, and mean interval width before and after Studentization. For a fixed marginal scale, the interval width ratio is $t_{n,0.975}/\Phi^{-1}(0.975)$, where $t_{n,0.975}$ is the 97.5th percentile of a univariate Student-t distribution with n degrees of freedom and Φ is the standard Gaussian distribution function. At the training sample sizes used here, Studentization increased interval widths by approximately 0.18% to 0.45%. Coverage therefore changed little in these experiments. For example, RCGP retained 100% empirical coverage in all eight uniform and asymmetric conditions, with wide intervals, while the mixture GP retained approximately 84% coverage on Yacht with focused contamination.

Table 12: RCGP: absolute predictive scores and intervals. Entries are means across ten splits. G and St denote Gaussian and Studentized predictions from the same fit. PLL is in nats per response, coverage is a percentage, and widths are in original response units.

Data	Contamination	Mean pointwise PLL		Joint PLL / p		95% coverage		Width	
		G	St	G	St	G	St	G	St
Synthetic	clean	0.800	0.799	0.806	0.805	96.3	96.3	0.419	0.421
	uniform	−1.562	−1.563	−1.560	−1.536	100.0	100.0	7.371	7.405
	asymmetric	−1.559	−1.560	−1.555	−1.533	100.0	100.0	7.186	7.218
	focused	−0.718	−0.698	−0.687	−0.676	95.7	95.7	1.978	1.987
Boston	clean	−2.385	−2.384	−2.371	−2.371	95.5	95.5	11.809	11.840
	uniform	−3.885	−3.885	−3.883	−3.860	100.0	100.0	74.279	74.477
	asymmetric	−3.864	−3.865	−3.860	−3.839	100.0	100.0	71.360	71.550
	focused	−2.977	−2.976	−2.966	−2.961	97.8	97.8	24.523	24.588
Energy	clean	−0.603	−0.601	−0.565	−0.564	93.8	93.8	1.774	1.777
	uniform	−3.916	−3.917	−3.915	−3.890	100.0	100.0	77.365	77.500
	asymmetric	−3.911	−3.911	−3.907	−3.885	100.0	100.0	75.111	75.242
	focused	−0.947	−0.947	−0.883	−0.876	98.1	98.1	3.814	3.821
Yacht	clean	0.086	0.088	0.091	0.093	95.5	95.5	1.229	1.235
	uniform	−4.337	−4.338	−4.334	−4.311	100.0	100.0	116.681	117.193
	asymmetric	−4.306	−4.307	−4.301	−4.280	100.0	100.0	111.822	112.312
	focused	−1.936	−1.935	−1.847	−1.829	99.4	99.4	15.064	15.130

Table 13: Relevance Pursuit: absolute predictive scores and intervals. Entries are means across ten splits. G and St denote Gaussian and Studentized predictions from the same fit. PLL is in nats per response, coverage is a percentage, and widths are in original response units.

Data	Contamination	Mean pointwise PLL		Joint PLL / p		95% coverage		Width	
		G	St	G	St	G	St	G	St
Synthetic	clean	0.786	0.785	0.787	0.786	95.3	95.3	0.423	0.425
	uniform	0.792	0.791	0.793	0.792	95.3	95.7	0.429	0.431
	asymmetric	0.794	0.793	0.793	0.793	95.0	95.0	0.424	0.425
	focused	−12.771	−6.342	−11.931	−5.054	87.7	88.7	0.465	0.467
Boston	clean	−2.364	−2.362	−2.348	−2.346	92.2	92.2	9.959	9.986
	uniform	−2.517	−2.504	−2.497	−2.477	92.2	92.2	10.014	10.041
	asymmetric	−2.650	−2.636	−2.615	−2.587	88.2	88.4	8.943	8.967
	focused	−2.439	−2.438	−2.426	−2.425	94.7	94.7	13.079	13.114
Energy	clean	−0.651	−0.650	−0.643	−0.642	94.4	94.4	1.939	1.942
	uniform	−0.673	−0.671	−0.660	−0.649	90.4	90.4	1.696	1.699
	asymmetric	−0.648	−0.644	−0.638	−0.622	91.8	91.8	1.618	1.621
	focused	−0.642	−0.641	−0.637	−0.636	94.0	94.0	1.906	1.910
Yacht	clean	−0.698	−0.619	−0.618	−0.558	90.3	90.3	1.042	1.047
	uniform	−0.578	−0.569	−0.526	−0.516	96.1	96.5	2.032	2.041
	asymmetric	−0.515	−0.508	−0.470	−0.459	97.1	97.1	1.987	1.995
	focused	−69.325	−5.710	−70.403	−5.945	95.8	95.8	2.291	2.301

Table 14: Mixture GP: absolute predictive scores and intervals. Entries are means across ten splits. G and St denote Gaussian and Studentized predictions from the same fit. PLL is in nats per response, coverage is a percentage, and widths are in original response units.

Data	Contamination	Mean pointwise PLL		Joint PLL / p		95% coverage		Width	
		G	St	G	St	G	St	G	St
Synthetic	clean	0.802	0.801	0.805	0.805	95.3	95.7	0.415	0.417
	uniform	0.802	0.802	0.806	0.805	95.7	95.7	0.422	0.423
	asymmetric	0.802	0.801	0.806	0.805	95.7	95.7	0.420	0.422
	focused	−14.262	−7.546	−12.376	−4.875	86.3	86.3	0.442	0.444
Boston	clean	−2.483	−2.474	−2.465	−2.454	90.0	90.0	8.889	8.913
	uniform	−2.779	−2.743	−2.760	−2.727	90.6	90.6	13.270	13.306
	asymmetric	−2.486	−2.482	−2.465	−2.462	92.2	92.2	11.202	11.232
	focused	−3.132	−3.122	−3.119	−3.111	92.0	92.2	18.754	18.804
Energy	clean	−0.620	−0.619	−0.587	−0.586	95.2	95.2	1.790	1.793
	uniform	−2.710	−2.268	−3.338	−2.437	91.3	91.3	3.295	3.301
	asymmetric	−2.083	−1.905	−3.093	−2.172	91.8	91.8	3.175	3.181
	focused	−0.990	−0.986	−0.933	−0.929	96.5	96.5	3.056	3.062
Yacht	clean	−0.016	−0.008	0.037	0.056	88.4	88.7	1.303	1.309
	uniform	−7.772	−1.560	−7.777	−1.654	90.6	90.6	1.250	1.255
	asymmetric	−0.459	−0.390	−0.445	−0.361	91.0	91.0	1.206	1.211
	focused	−160.431	−7.594	−167.688	−7.432	83.5	83.5	2.880	2.893

G.5 Prediction blocks and the original data processing

To check sensitivity to prediction dimension, we also evaluated joint PLL on nonoverlapping blocks of $p = 16$ test observations, divided each score by p, and averaged across complete blocks within each split. Under this evaluation, Studentization improved joint PLL averaged within each condition in 15/16 conditions for RCGP and 13/16 for each of the other two models. Blocks were formed using the shared random ordering of test observations. Incomplete final blocks were omitted only from this supplementary evaluation. The main joint PLL used every held out observation.

We also evaluated one fixed split per condition using the released RCGP data processing conventions. These conventions use the full dataset to standardize the real data and 20% contamination for the synthetic asymmetric condition. Under these conventions, joint PLL improved in 13/16 conditions for RCGP, 12/16 for Relevance Pursuit, and 9/16 for the mixture GP. The corresponding pointwise counts were 4/16, 7/16, and 9/16. These results are reported separately in Tables 15, 16, and 17 because both the normalization and the use of one fixed split per condition differ from the main evaluation.

Table 15: RCGP: absolute predictive scores and intervals. Each condition uses one fixed split with the original data processing; synthetic asymmetric contamination is 20%. G and St denote Gaussian and Studentized predictions from the same fit. PLL is in nats per response, coverage is a percentage, and widths are in original response units.

Data	Contamination	Mean pointwise PLL		Joint PLL / p		95% coverage		Width	
		G	St	G	St	G	St	G	St
Synthetic	clean	0.801	0.800	0.808	0.806	96.7	96.7	0.417	0.419
	uniform	−1.574	−1.575	−1.573	−1.548	100.0	100.0	7.497	7.530
	asymmetric	−1.938	−1.939	−1.932	−1.913	100.0	100.0	10.148	10.194
	focused	−0.457	−0.457	−0.453	−0.443	96.7	96.7	2.092	2.102
Boston	clean	−2.347	−2.346	−2.347	−2.348	94.1	94.1	10.476	10.504
	uniform	−3.907	−3.907	−3.906	−3.881	100.0	100.0	76.675	76.880
	asymmetric	−3.852	−3.853	−3.848	−3.826	100.0	100.0	70.872	71.061
	focused	−3.056	−3.055	−3.043	−3.041	96.1	96.1	24.315	24.380
Energy	clean	−0.418	−0.418	−0.400	−0.396	98.7	98.7	1.863	1.867
	uniform	−3.936	−3.936	−3.935	−3.910	100.0	100.0	79.215	79.354
	asymmetric	−3.882	−3.883	−3.881	−3.857	100.0	100.0	74.152	74.282
	focused	−0.981	−0.981	−0.912	−0.899	98.7	98.7	3.970	3.977
Yacht	clean	0.092	0.091	0.128	0.129	100.0	100.0	1.725	1.732
	uniform	−4.358	−4.359	−4.356	−4.333	100.0	100.0	119.260	119.783
	asymmetric	−4.291	−4.292	−4.288	−4.266	100.0	100.0	110.255	110.739
	focused	−3.050	−3.037	−3.018	−3.019	96.8	96.8	25.917	26.031

G.6 Optimization diagnostics

All 528 model evaluations produced finite predictive scores, including the 480 main evaluations and 48 evaluations using the original data processing. The saved RCGP convergence flag was positive for 173 of 176 fits, while the final optimizer call at the selected Relevance Pursuit support reported success for 167 of 176 fits. Only three Relevance Pursuit fits reported success at every support explored during the search. The mixture GP used a fixed update budget, which does not establish convergence. All finite fits were retained, including those that reached the maximum number of iterations.

G.7 Proof of Proposition 1 and numerical bounds

The bound in Proposition 1 follows directly from the density ratio of the two predictive distributions. For a positive definite fitted scale $\mathbf{\Sigma}$, write $r^2 = (\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ and let p be the prediction

Table 16: Relevance Pursuit: absolute predictive scores and intervals. Each condition uses one fixed split with the original data processing; synthetic asymmetric contamination is 20%. G and St denote Gaussian and Studentized predictions from the same fit. PLL is in nats per response, coverage is a percentage, and widths are in original response units.

Data	Contamination	Mean pointwise PLL		Joint PLL / p		95% coverage		Width	
		G	St	G	St	G	St	G	St
Synthetic	clean	0.784	0.783	0.785	0.783	100.0	100.0	0.421	0.423
	uniform	0.806	0.805	0.807	0.805	100.0	100.0	0.427	0.429
	asymmetric	0.777	0.776	0.781	0.779	100.0	100.0	0.429	0.430
	focused	−7.388	−3.927	−7.385	−4.084	96.7	96.7	0.452	0.454
Boston	clean	−2.347	−2.342	−2.342	−2.342	94.1	94.1	9.018	9.042
	uniform	−2.791	−2.777	−2.763	−2.744	92.2	92.2	9.752	9.779
	asymmetric	−2.386	−2.381	−2.375	−2.375	94.1	94.1	9.413	9.438
	focused	−2.398	−2.396	−2.386	−2.386	94.1	94.1	12.244	12.276
Energy	clean	−0.580	−0.580	−0.570	−0.568	94.8	94.8	1.975	1.978
	uniform	−0.468	−0.469	−0.462	−0.458	96.1	96.1	1.897	1.900
	asymmetric	−0.483	−0.483	−0.469	−0.464	96.1	96.1	1.950	1.953
	focused	−0.587	−0.588	−0.579	−0.578	96.1	97.4	1.971	1.974
Yacht	clean	−0.414	−0.388	−0.456	−0.444	90.3	90.3	1.095	1.100
	uniform	−0.372	−0.373	−0.370	−0.356	100.0	100.0	2.090	2.099
	asymmetric	−0.328	−0.329	−0.325	−0.311	100.0	100.0	2.030	2.039
	focused	−393.994	−21.238	−394.655	−22.546	93.5	93.5	3.024	3.037

Table 17: Mixture GP: absolute predictive scores and intervals. Each condition uses one fixed split with the original data processing; synthetic asymmetric contamination is 20%. G and St denote Gaussian and Studentized predictions from the same fit. PLL is in nats per response, coverage is a percentage, and widths are in original response units.

Data	Contamination	Mean pointwise PLL		Joint PLL / p		95% coverage		Width	
		G	St	G	St	G	St	G	St
Synthetic	clean	0.806	0.805	0.810	0.809	96.7	100.0	0.414	0.416
	uniform	0.816	0.815	0.821	0.819	96.7	96.7	0.421	0.422
	asymmetric	0.781	0.781	0.787	0.786	96.7	96.7	0.415	0.417
	focused	−8.141	−4.721	−8.005	−4.298	86.7	86.7	0.441	0.443
Boston	clean	−2.185	−2.181	−2.180	−2.181	98.0	98.0	8.305	8.327
	uniform	−2.237	−2.235	−2.219	−2.220	94.1	94.1	8.778	8.801
	asymmetric	−2.238	−2.235	−2.219	−2.220	94.1	94.1	8.692	8.716
	focused	−2.213	−2.211	−2.199	−2.200	96.1	96.1	8.558	8.581
Energy	clean	−0.455	−0.455	−0.430	−0.427	98.7	98.7	1.859	1.862
	uniform	−0.522	−0.522	−0.500	−0.495	98.7	98.7	2.085	2.089
	asymmetric	−0.518	−0.518	−0.496	−0.491	98.7	98.7	2.069	2.073
	focused	−0.884	−0.884	−0.821	−0.818	96.1	96.1	2.869	2.874
Yacht	clean	0.110	0.113	0.075	0.079	93.5	93.5	1.217	1.222
	uniform	0.108	0.108	0.144	0.145	96.8	96.8	1.514	1.521
	asymmetric	0.064	0.066	0.060	0.061	90.3	90.3	1.324	1.330
	focused	0.043	0.045	0.029	0.031	90.3	90.3	1.331	1.337

dimension. The change in joint PLL is

$$\ell(r^2) = \log \frac{\mathrm{St}_n(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}$$
$$= \log \frac{\Gamma((n+p)/2)}{\Gamma(n/2)} - \frac{p}{2} \log \frac{n}{2} + \frac{r^2}{2} - \frac{n+p}{2} \log \left(1 + \frac{r^2}{n}\right).$$

The derivative with respect to r^2 is $(r^2 - p)/(2(n + r^2))$. The minimum therefore occurs at $r^2 = p$. Consequently, $\ell(r^2)/p \geq \ell(p)/p = -\varepsilon_{n,p}$ for every response vector and every fitted mean and positive definite scale. Each Student-t marginal has n degrees of freedom and the same marginal scale as its Gaussian counterpart. Applying the $p = 1$ bound to each marginal and averaging gives the lower bound $-\varepsilon_{n,1}$ for mean pointwise PLL, which completes the proof. □

The bounds for the present experiments are shown in Table 18, rounded downward to preserve the inequality. These bounds apply to individual predictions and hence also to the split and condition averages used here. They bound the possible loss from this correction without assuming that the fitted model is correct.

Table 18: Lower bounds on the change in PLL from Studentization with the predictive scale held fixed, in 10^{-3} nats per response. Values are rounded downward to preserve the inequality.

Data	Mean pointwise PLL	Joint PLL / p
Synthetic	−1.851	−1.759
Boston	−1.099	−1.043
Energy	−0.724	−0.687
Yacht	−1.804	−1.713