1 The GRwHS Model

1.1 Notations and Problem Setup

Consider a dataset with n observations and p features. Each observation consists of a feature vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$ and a continuous response $y_i \in \mathbb{R}$. The data are collected into an $n \times p$ design matrix $\boldsymbol{X}$ and a response vector $\boldsymbol{y}$. We assume the standard linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n).$$

To make the hyperpriors globally interpretable and to stabilize computation, we assume that the response $\boldsymbol{y}$ is centered and that each column of $\boldsymbol{X}$ has been standardized to *zero mean and unit variance.*[1]

Furthermore, the p features are partitioned into G disjoint groups, $\mathcal{G}_1, \ldots, \mathcal{G}_G$. Our goal is to estimate the regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ by leveraging this group structure, capturing mixed-signal scenarios where a few strong effects coexist with numerous weak but informative signals.

1.2 Hierarchical Prior for Mixed Signals

To handle the mixed-signal challenge, we introduce a hierarchical prior that separates intra-group adaptive selection from inter-group stabilization via a global–local shrinkage design with multiplicative scales. The resulting prior for a coefficient β_j in group $\mathcal{G}_g$ is:

$$\beta_j \mid \phi_g, \tau, \lambda_j, \sigma^2 \sim \mathcal{N}(0, \phi_g^2 \tau^2 \tilde{\lambda}_j^2 \sigma^2).$$

This structure is composed of two components detailed below.

1.2.1 Intra-Group Selection via Regularized Horseshoe

For adaptive sparse selection within each group, we employ the Regularized Horseshoe (RHS) prior Carvalho et al., 2010; Piironen and Vehtari, 2017. This introduces a local scale λ_j for each coefficient and a global scale τ controlling overall sparsity:

$$\begin{aligned} \lambda_j &\sim \mathrm{C}^+(0,1), \quad j = 1, \ldots, p, \\ \tau &\sim \mathrm{C}^+(0, \tau_0). \end{aligned}$$

Here $\mathrm{C}^+(0, \cdot)$ denotes a Half-Cauchy prior with density $p(\theta \mid s) = \dfrac{2s}{\pi(s^2 + \theta^2)}$ for $\theta > 0$ and scale s. The regularized local variance is defined as $\tilde{\lambda}_j^2 = \frac{c^2 \lambda_j^2}{c^2 + \tau^2 \lambda_j^2}$, which introduces a slab-width parameter c to stabilize heavy tails. With standardized data, the global scale can be set using the effective sparsity heuristic

$$\tau_0 \approx \frac{s}{p-s} \frac{1}{\sqrt{n}},$$

where s is a prior guess of relevant coefficients. If instead the predictors are scaled to unit ℓ_2-norm per column, use $\tau_0 \approx \frac{s}{p-s}$ (i.e., multiply the above by $\sqrt{n}$). The slab width c can be chosen based on the expected magnitude of non-zero effects, e.g., $c \in [1, 2]$ in units of σ.

[1] Some implementations instead rescale each column to have unit ℓ_2-norm, i.e., $\|\boldsymbol{x}_{\cdot j}\|_2 = 1$, which implies $\mathrm{sd}(x_{\cdot j}) \approx 1/\sqrt{n}$. If that convention is used, the global-sparsity heuristic for τ_0 below should be multiplied by $\sqrt{n}$ to maintain the same prior scale. For unit-variance columns, use $\tau_0 \approx \frac{s}{p-s}\frac{1}{\sqrt{n}}$; for unit-$\ell_2$ columns, use $\tau_0 \approx \frac{s}{p-s}$.

1.2.2 Inter-Group Stabilization via Size-Adjusted Half-Normal

To account for weak signals dispersed across groups without favoring small groups, each group $\mathcal{G}_g$ has a group-level scale ϕ_g with a size-adjusted Half-Normal prior:

$$\phi_g \sim \mathcal{N}^+\big(0, \eta_g^2\big), \qquad \eta_g \equiv \frac{\eta}{\sqrt{p_g}}, \quad g = 1, \ldots, G.$$

Here $\mathcal{N}^+(0, \cdot)$ denotes a Half-Normal prior. This provides stabilizing ridge-like shrinkage by sharing information within group g, thereby mitigating over-shrinkage of weak signals while allowing strong effects to remain expressive. Marginally, while β_j inherits heavy tails from λ_j, it also gains within-group variance pooling through the common scale ϕ_g. For standardized predictors, a sensible choice for the base scale is $\eta \in [0.3, 1]$; larger values weaken pooling.[2]

The combination of these components yields a **group-regularized horseshoe**. Unlike a pure group-horseshoe that places heavy-tailed priors on the group scales ϕ_g, our Half-Normal choice trades heavier group-level tails for more stability, making the model well-suited to problems where many weak signals are dispersed across groups.

1.3 Interpreting Shrinkage and Variance Components

Let d_j denote the prior precision on β_j, $d_j = (\phi_{g(j)}^2 \tau^2 \tilde{\lambda}_j^2 \sigma^2)^{-1}$, and $q_j = \sigma^{-2}(X^\top X)_{jj}$. A ridge-style diagonal proxy for the per-coordinate shrinkage is

$$\kappa_j \;=\; \frac{q_j}{q_j + d_j} \in (0, 1),$$

interpretable as the fraction of OLS signal preserved in the posterior mean (orthonormal design: $\kappa_j = \frac{n}{n+\sigma^2 d_j}$).[3]

A normalized variance budget (priors-only). To attribute shrinkage across hierarchical *prior* components in a way that sums to one and is invariant to the noise scale, define raw log-weights

$$a_{g(j)} = 2 \log \phi_{g(j)}, \quad a_\tau = 2 \log \tau, \quad a_{\lambda_j} = 2 \log \tilde{\lambda}_j.$$

For numerical stability we apply a symmetric clipping $a_k^\delta = \operatorname{sign}(a_k) \cdot \max(|a_k|, \delta)$ with small $\delta > 0$ (e.g. 10^{-8}), and then set the *normalized* diagnostics

$$\omega_{g(j)} = \frac{a_{g(j)}^\delta}{a_{g(j)}^\delta + a_\tau^\delta + a_{\lambda_j}^\delta}, \qquad \omega_\tau = \frac{a_\tau^\delta}{a_{g(j)}^\delta + a_\tau^\delta + a_{\lambda_j}^\delta}, \qquad \omega_{\lambda_j} = \frac{a_{\lambda_j}^\delta}{a_{g(j)}^\delta + a_\tau^\delta + a_{\lambda_j}^\delta},$$

so that $\omega_{g(j)} + \omega_\tau + \omega_{\lambda_j} = 1$ by construction.[4]

[2] Compared with a Half-Cauchy on ϕ_g, the Half-Normal has lighter tails. This deliberate choice enhances stability for scenarios with many weak, dispersed signals; one may assess robustness by sensitivity checks over η. If an unadjusted variant is preferred, set $\eta_g \equiv \eta$.

[3] **Caveat for correlated designs:** the proxy based on $(X^\top X)_{jj}$ can mislead when predictors are correlated. For validation, one can additionally estimate the diagonal of the hat matrix $H = X(X^\top X + D_\beta \sigma^2)^{-1} X^\top$ via Hutchinson+CG; see Algorithm 4.

[4] If one wishes to include the noise scale in the budget, add $a_\sigma = 2 \log \sigma$ and renormalize over four terms.

2 Bayesian Computation

2.1 Exact Inference via Gibbs Sampling

We employ a blocked Gibbs scheme with auxiliary variables for Half-Cauchy priors[5] Makalic and Schmidt, 2016 and slice sampling steps where conjugacy does not hold. The multiplicative scales can potentially lead to slow MCMC mixing, which we partly mitigate by sampling scale parameters on the log-transform.

MCMC Stability and Reparameterization. The prior variance of coefficient β_j in group $g = g(j)$ is

$$(\beta_j \mid \phi_g, \tau, \lambda_j, \sigma) = \phi_g^2\, \tau^2\, \tilde{\lambda}_j^2\, \sigma^2, \qquad \tilde{\lambda}_j^2 = \frac{c^2 \lambda_j^2}{c^2 + \tau^2 \lambda_j^2}.$$

Two limiting regimes induce near-nonidentifiability among $(\phi_g, \tau, \lambda_j)$: *(i) Spike regime* $\tau^2\lambda_j^2 \ll c^2$, where $\tilde{\lambda}_j^2 \approx \lambda_j^2$ and the product $\phi_g \tau \lambda_j$ alone controls the scale; and *(ii) Slab regime* $\tau^2\lambda_j^2 \gg c^2$, where $\tilde{\lambda}_j^2 \approx c^2/\tau^2$ hence $(\beta_j|\cdot) \approx \phi_g^2 c^2 \sigma^2$ becomes nearly insensitive to (λ_j, τ). These create strong posterior dependencies and impede mixing.

To address this, we use a non-centered log-scale reparameterization. Define log-scales $z_g = \log \phi_g$, $u_j = \log \lambda_j$, $v = \log \tau$ and a standard Gaussian $\epsilon_j \sim \mathcal{N}(0, 1)$. We reparameterize

$$\beta_j \;=\; \sigma\, \exp(z_{g(j)} + v)\, \tilde{\lambda}_j(u_j, v)\, \epsilon_j, \qquad \tilde{\lambda}_j(u_j, v) = \frac{c\, \exp(u_j)}{\sqrt{c^2 + \exp(2u_j + 2v)}}.$$

Sampling in (z_g, u_j, v) stabilizes the geometry and makes scale trade-offs explicit in log-space. We update the transformed variables with slice/MH moves on (z_g, u_j, v), and sample ϵ_j from its standard Normal.

Anchoring for interpretability (post-processing only). To avoid arbitrary drift across layers in summaries, we adopt a post-processing anchoring: report $z_g^\star = z_g - \overline{z}$ with $\overline{z} = \frac{1}{G}\sum_g z_g$, and leave v unchanged. This does not modify the MCMC or VI updates and therefore does not alter the posterior; it is purely a reporting convention that makes the relative contributions of $(\phi_g, \tau, \lambda_j)$ comparable across runs.

Gaussian block for $\boldsymbol{\beta}$ (scaled Woodbury). Introduce $\boldsymbol{D}_\beta = \operatorname{diag}(d_{jj})$ with $d_{jj} = (\phi_{g(j)}^2 \tau^2 \tilde{\lambda}_j^2 \sigma^2)^{-1}$. Then

$$\boldsymbol{\beta} \mid \tau, \boldsymbol{\lambda}, \boldsymbol{\phi}, \sigma^2, \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta),$$

with

$$\begin{aligned}
\boldsymbol{\Sigma}_\beta^{-1} &= \sigma^{-2}\Big(\boldsymbol{X}^T\boldsymbol{X} + \operatorname{diag}\big((\phi_{g(j)}^2 \tau^2 \tilde{\lambda}_j^2)^{-1}\big)\Big), \\
\boldsymbol{\mu}_\beta &= \boldsymbol{\Sigma}_\beta(\sigma^{-2}\boldsymbol{X}^T\boldsymbol{y}).
\end{aligned}$$

Working with the σ^{-2}-scaled precision reduces dynamic-range issues. For $p > n$, compute $\boldsymbol{\Sigma}_\beta$ via the Woodbury identity on the scaled system:

$$\boldsymbol{\Sigma}_\beta = \Big(\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{D}_\beta\Big)^{-1} = \boldsymbol{D}_\beta^{-1} - \boldsymbol{D}_\beta^{-1}\boldsymbol{X}^T\Big(\sigma^2\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{D}_\beta^{-1}\boldsymbol{X}^T\Big)^{-1}\boldsymbol{X}\boldsymbol{D}_\beta^{-1}.$$

[5] We use the shape–scale parameterization for the Inverse Gamma distribution, InvGamma(α, β), with density proportional to $\beta^\alpha x^{-\alpha-1}e^{-\beta/x}$.

Auxiliary variables for Half-Cauchy scales. To aid computation, we use the equivalent inverse-gamma mixture representation for the Half-Cauchy priors, adopting the shape–scale parameterization for $\mathrm{InvGamma}(\alpha, \beta)$. The full hierarchy for the local, global, and noise scales is:

$$\begin{aligned}
\lambda_j^2 \mid \xi_j &\sim \mathrm{InvGamma}\Big(\tfrac{1}{2}, \tfrac{1}{\xi_j}\Big), & \xi_j &\sim \mathrm{InvGamma}\big(\tfrac{1}{2}, 1\big) & &\text{(Local scales)}\\
\tau^2 \mid \xi_\tau &\sim \mathrm{InvGamma}\Big(\tfrac{1}{2}, \tfrac{1}{\xi_\tau}\Big), & \xi_\tau &\sim \mathrm{InvGamma}\Big(\tfrac{1}{2}, \tfrac{1}{\tau_0^2}\Big) & &\text{(Global scale)}\\
\sigma^2 \mid \xi_\sigma &\sim \mathrm{InvGamma}\Big(\tfrac{1}{2}, \tfrac{1}{\xi_\sigma}\Big), & \xi_\sigma &\sim \mathrm{InvGamma}\Big(\tfrac{1}{2}, \tfrac{1}{s_0^2}\Big) & &\text{(Noise scale)}
\end{aligned}$$

While this representation does not restore full conjugacy for λ_j and τ under the RHS prior ($c < \infty$), it yields simple conjugate forms for the conditional posteriors of the auxiliary variables. These are essential steps in the Gibbs sampler:

$$\begin{aligned}
\xi_j \mid \lambda_j^2 &\sim \mathrm{InvGamma}\Big(1,\ 1 + \tfrac{1}{\lambda_j^2}\Big),\\
\xi_\tau \mid \tau^2 &\sim \mathrm{InvGamma}\Big(1,\ \tfrac{1}{\tau_0^2} + \tfrac{1}{\tau^2}\Big),\\
\xi_\sigma \mid \sigma^2 &\sim \mathrm{InvGamma}\Big(1,\ \tfrac{1}{s_0^2} + \tfrac{1}{\sigma^2}\Big).
\end{aligned}$$

Local scales λ_j. Because $\tilde{\lambda}_j^2 = \frac{c^2\lambda_j^2}{c^2+\tau^2\lambda_j^2}$ couples λ_j with τ, the conditional posterior is not conjugate. We update each λ_j via slice sampling (or Metropolis–Hastings) on $u_j = \log\lambda_j$ with log-target proportional to

$$-\log\tilde{\lambda}_j - \frac{\beta_j^2}{2\phi_{g(j)}^2\tau^2\tilde{\lambda}_j^2\sigma^2} + \log \mathrm{C}^+(\lambda_j \mid 0, 1) + u_j,$$

where the final $+u_j$ term is the Jacobian for the log-reparameterization. Optionally refresh ξ_j from its inverse-gamma conditional given λ_j^2 as above.

Group scales ϕ_g. Let $\theta_g = \phi_g^2$. Since $\phi_g \sim \mathcal{N}^+(0, \eta_g^2)$ with $\eta_g = \eta/\sqrt{p_g}$ induces a density proportional to $\theta_g^{-1/2}\exp(-\theta_g/(2\eta_g^2))$ over $\theta_g > 0$, combining with the Gaussian prior on β_j yields

$$p(\theta_g \mid \boldsymbol{\beta}_{\mathcal{G}_g}, \tau, \boldsymbol{\lambda}, \sigma^2) \propto \theta_g^{-(p_g+1)/2}\exp\left(-\frac{1}{2\tau^2\sigma^2\theta_g}\sum_{j\in\mathcal{G}_g}\frac{\beta_j^2}{\tilde{\lambda}_j^2} - \frac{\theta_g}{2\eta_g^2}\right),$$

which is a generalized inverse Gaussian distribution:

$$\theta_g \mid \cdots \sim \mathrm{GIG}\left(\tfrac{1}{2} - \tfrac{p_g}{2},\ \frac{1}{\tau^2\sigma^2}\sum_{j\in\mathcal{G}_g}\frac{\beta_j^2}{\tilde{\lambda}_j^2},\ \frac{1}{\eta_g^2}\right), \qquad \phi_g = \sqrt{\theta_g}.$$

Global scale τ. Under RHS ($c < \infty$), $\tilde{\lambda}_j^2$ depends on τ, therefore the conditional posterior of τ is not inverse-gamma. We update $v = \log\tau$ via slice sampling (or Metropolis–Hastings) with a log-target that combines multiple terms:

$$\underbrace{-\,p\,\log\tau - \sum_{j=1}^{p}\log\tilde{\lambda}_j}_{\text{from Normalizer of } p(\boldsymbol{\beta})}\ \underbrace{-\frac{1}{2\sigma^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\phi_{g(j)}^2\tau^2\tilde{\lambda}_j^2}}_{\text{from Likelihood of } \boldsymbol{\beta}}\ \underbrace{+\log \mathrm{C}^+(\tau \mid 0, \tau_0)}_{\text{Prior on } \tau}\ \underbrace{+v}_{\text{Jacobian}}$$

where $\tilde{\lambda}_j^2 = \dfrac{c^2\lambda_j^2}{c^2 + \tau^2\lambda_j^2}$ and the final $+v$ term is the Jacobian for the log-reparameterization. Optionally refresh ξ_τ from its conditional given τ^2 as above.

Noise variance σ^2. Using the auxiliary variable ξ_σ for a Half-Cauchy prior on σ with scale s_0, the conditional posterior for σ^2 is an Inverse Gamma distribution:

$$\begin{aligned}
\sigma^2 \mid \ldots &\sim \mathrm{InvGamma}(\alpha, \beta), \quad \text{where} \\
\alpha &= \frac{n+p+1}{2}, \\
\beta &= \underbrace{\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2}_{\text{Likelihood (RSS)}} + \underbrace{\frac{1}{2}\sum_{j=1}^{p} \frac{\beta_j^2}{\phi_{g(j)}^2 \tau^2 \tilde{\lambda}_j^2}}_{\text{Prior on } \boldsymbol{\beta}} + \underbrace{\frac{1}{\xi_\sigma}}_{\text{Prior on } \sigma} \quad .
\end{aligned}$$

The shape parameter α combines contributions from the likelihood ($n/2$), the prior on $\boldsymbol{\beta}$ ($p/2$), and the auxiliary-variable representation of the Half-Cauchy prior on σ ($1/2$). The scale parameter β aggregates the corresponding variance components.

2.2 Approximate Inference via Structured Variational Inference

For scalable inference, especially in high-dimensional settings, we develop a structured variational inference (SVI) scheme. We follow the original exposition structure but augment it with a low-rank coupling between $(\log \tau)$ and $(\log \lambda_j)$, expanded natural-gradient details, and variance/complexity analysis.

2.2.1 The Variational Family (with low-rank coupling)

We preserve the within-group dependence between $\boldsymbol{\beta}_{\mathcal{G}_g}$ and ϕ_g, and introduce a one-factor coupling for $(\log \tau, \log \lambda_j)$:

$$q(\boldsymbol{\beta}, \boldsymbol{\lambda}, \tau, \boldsymbol{\phi}, \sigma^2) = \left[\prod_{g=1}^{G} q(\boldsymbol{\beta}_{\mathcal{G}_g}, \phi_g)\right] q(\boldsymbol{\lambda}, \tau)\, q(\sigma^2).$$

Parameterization:

$$\begin{gathered}
q(\boldsymbol{\beta}_{\mathcal{G}_g}, \log \phi_g) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m}_g \\ \mu_{\log \phi_g} \end{bmatrix}, \begin{bmatrix} \boldsymbol{S}_g & \boldsymbol{c}_g \\ \boldsymbol{c}_g^\top & v_{\log \phi_g} \end{bmatrix}\right), \\
\log \sigma \sim \mathcal{N}(\mu_{\log \sigma}, v_{\log \sigma}).
\end{gathered}$$

For the global–local block, let $u \sim \mathcal{N}(0,1)$ be a shared factor and set

$$\log \lambda_j = \mu_{\log \lambda_j} + a_j\, u + \epsilon_j, \quad \epsilon_j \sim \mathcal{N}(0, v_{\log \lambda_j}), \qquad \log \tau = \mu_{\log \tau} + \rho\, u + \epsilon_\tau, \quad \epsilon_\tau \sim \mathcal{N}(0, v_{\log \tau}).$$

When $\rho = a_j = 0$ we recover the original independent log-normal factors.

2.2.2 ELBO Roadmap and Notation

We split the expected log prior on $\boldsymbol{\beta}$ into a *group block* and a *global–local block* using identity (2.1). Moments under q use the shorthand

$$m_{\beta_j} = \mathbb{E}_q[\beta_j], \quad v_{\beta_j} = \mathrm{Var}_q(\beta_j), \quad \mu_{W_g} = \mathbb{E}_q[W_g], \quad s^2_{W_g} = \mathrm{Var}_q(W_g), \quad \Sigma_{\beta_j, W_g} = \mathrm{Cov}_q(\beta_j, W_g),$$

with $W_g = 2\log\phi_g + 2\log\sigma$. Under our factorization, $\log\sigma$ is independent of $(\boldsymbol{\beta}_{\mathcal{G}_g}, \log\phi_g)$, hence $\Sigma_{\beta_j, W_g} = 2\,\Sigma_{\beta_j, \log\phi_g}$.

2.2.3 Decomposition of the Evidence Lower Bound (ELBO)

The ELBO is

$$\mathcal{L}(q) = \underbrace{\mathbb{E}_q[\log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)]}_{\text{Expected Log-Likelihood}} + \underbrace{\mathbb{E}_q[\log p(\boldsymbol{\beta} \mid \boldsymbol{\lambda}, \tau, \boldsymbol{\phi}, \sigma^2)]}_{\text{Expected Log-Prior on } \beta} + \underbrace{\sum \mathbb{E}_q[\log p(\text{scales})]}_{\text{Expected Log-Priors on Scales}} + \mathbb{H}[q].$$

Expected log-likelihood. Because $q(\beta)$ and $q(\sigma)$ are independent,

$$\mathbb{E}_q[\log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)] = -\frac{n}{2}\,\mathbb{E}_q[\log(2\pi\sigma^2)] - \frac{1}{2}\,\mathbb{E}_q\Big[\frac{1}{\sigma^2}\Big]\,\Big(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{m}\|_2^2 + \mathrm{tr}\big(\boldsymbol{X}\ \mathrm{blkdiag}(\boldsymbol{S}_1, \ldots, \boldsymbol{S}_G)\,\boldsymbol{X}^\top\big)\Big).$$

RHS inverse-scale decomposition. For $c, \tau, \lambda > 0$ and $\tilde{\lambda}^2 = \frac{c^2\lambda^2}{c^2 + \tau^2\lambda^2}$,

$$\frac{1}{\tau^2\tilde{\lambda}^2} \;=\; \frac{1}{\tau^2\lambda^2} + \frac{1}{c^2}. \tag{2.1}$$

This identity will be used repeatedly in the ELBO derivation.

Lemma 2.1 (Gaussian–Exponential Quadratic Form). *If (β, W) is jointly Gaussian with mean (m_β, μ_W), variances (v_β, s_W^2), and covariance $\Sigma_{\beta W}$, then*

$$\mathbb{E}[\beta^2 e^{-W}] = \exp\big(-\mu_W + \tfrac{1}{2}s_W^2\big)\ \big(v_\beta + (m_\beta - \Sigma_{\beta W})^2\big).$$

Proof. See Appendix A. □

Group block. Taking $W_g = 2\log\phi_g + 2\log\sigma$, Lemma 2.1 yields

$$\mathbb{E}_q\Big[\frac{\beta_j^2}{\phi^2_{g(j)}\sigma^2}\Big] = \exp\Big(-\mu_{W_g} + \tfrac{1}{2}s^2_{W_g}\Big)\Big(v_{\beta_j} + (m_{\beta_j} - \Sigma_{\beta_j, W_g})^2\Big),$$

and, under our factorization, $\Sigma_{\beta_j, W_g} = 2\,\Sigma_{\beta_j, \log\phi_g}$.

Global–local block with coupling. Let $U_j = -2(\log\tau + \log\lambda_j)$. Under the one-factor coupling,

$$U_j \sim \mathcal{N}\big(\mu_{U_j}, v_{U_j}\big), \quad \mu_{U_j} = -2(\mu_{\log\tau} + \mu_{\log\lambda_j}), \quad v_{U_j} = 4\big(v_{\log\tau} + v_{\log\lambda_j} + (\rho + a_j)^2\big).$$

Using the log-normal identities above,

$$\mathbb{E}_q\Big[\frac{1}{\tau^2\lambda_j^2}\Big] = \exp\Big(-2(\mu_{\log\tau} + \mu_{\log\lambda_j}) + 2\big(v_{\log\tau} + v_{\log\lambda_j} + (\rho + a_j)^2\big)\Big).$$

Scale-prior terms (low-variance MC). Terms like $\mathbb{E}_q[\log \mathrm{C}^+(\tau \mid 0, \tau_0)]$ do not have a closed form under a log-normal q. We estimate these using a small number of Monte Carlo samples $L \in [8, 16]$ with a control variate given by a local quadratic expansion of the log-density in log-space; this has negligible cost relative to linear algebra with $\boldsymbol{X}$.

2.2.4 Natural-Gradient Details

Group block $(\boldsymbol{\beta}_{\mathcal{G}_g}, \log \phi_g)$**.** Write $\boldsymbol{\theta}_g = (\boldsymbol{\beta}_{\mathcal{G}_g}, \log \phi_g)$. Parameterize $q(\boldsymbol{\theta}_g)$ by Gaussian natural parameters $(\eta_{1,g}, \eta_{2,g})$ with $\eta_{1,g} = (S_g^{\theta})^{-1} m_g^{\theta}$ and $\eta_{2,g} = -\frac{1}{2}(S_g^{\theta})^{-1}$. The natural gradient matches sufficient statistics between the current q and the tilted distribution formed by the likelihood and prior contributions that involve $\boldsymbol{\theta}_g$. All expectations are analytic except matvecs with $X_{\mathcal{G}_g}/X_{\mathcal{G}_g}^{\top}$, handled with mini-batches.

Global–local block $(\log \lambda, \log \tau)$ **with one-factor coupling.** Let $\boldsymbol{\zeta} = (\log \lambda_1, \ldots, \log \lambda_p, \log \tau)$ be affine in (u, ϵ) as specified. The Fisher structure for the Gaussian latent-affine model decouples mean and covariance updates. The natural gradients for (a_j, ρ) equate the model-implied covariance $\mathrm{Cov}_q(\log \lambda_j, \log \tau)$ with its tilted counterpart; we clip (a_j, ρ) early-on to avoid over-coupling. Gradients of the Half-Cauchy log-priors are estimated with the same MC/control-variate samples used for the ELBO.

Noise block $\log \sigma$**.** Closed-form contribution from the likelihood plus MC/control-variates for the Half-Cauchy prior; we recommend a modestly smaller stepsize for $(\mu_{\log \sigma}, v_{\log \sigma})$ to prevent oscillations.

2.3 Computational Efficiency Enhancements

Practical defaults. Unless otherwise noted we recommend:

$m_{\text{Hutch}} = 10\text{–}20$, $\quad$ `cg_tol` $= 10^{-3}$ (or 10^{-4}), $\quad L_{\text{MC}} = 8\text{–}16$, $\quad c \in [1, 2]$, $\eta \in [0.3, 1]$, τ_0 as per scaling rule.

Algorithm 1 GRwHS inference overview

```
Require: Data $(\boldsymbol{X}, \boldsymbol{y})$; groups $\{\mathcal{G}_g\}$; hyperparameters $(c, \tau_0, \eta, s_0)$; choose $\textsc{Mode} \in \{\text{Gibbs}, \text{SVI}\}$.
Initialize coefficients/scales; standardize $X, y$; precompute $X^\top X$ if feasible.
if $\textsc{Mode} = \text{Gibbs}$ then
    for $t = 1{:}T$ do
        Sample $\boldsymbol{\beta} \mid \cdot$ via scaled precision and Woodbury when $p > n$.
        Update local $u_j = \log \lambda_j$ by slice/MH under RHS; optional IG auxiliaries.
        Update group $\theta_g = \phi_g^2$ via $\text{GIG}(\cdot)$; robust RoU when $\lambda \leq 0$.
        Update global $v = \log \tau$ by slice/MH; optional IG auxiliary.
        Update noise $\sigma^2$ by $\text{InvGamma}(\cdot)$ under Half-Cauchy mixture.
    end for
else                                                                    ▷ SVI
    for $t = 1{:}T$ do
        Draw minibatch; sample reparameterized variational latents with one-factor coupling
            $(\log \lambda, \log \tau)$; group block $(\boldsymbol{\beta}_{\mathcal{G}_g}, \log \phi_g)$ Gaussian.
        Build ELBO: analytic likelihood/group terms; low-variance MC for Half-Cauchy logs.
        Natural gradients; damping for covariances; early clipping for $(a_j, \rho)$.
        Update variational parameters; optional Polyak averaging for reporting.
    end for
end if
Diagnostics (post-processing): compute hat-diagonal via Hutchinson+CG; report $\kappa_j$,
    group edf, and normalized budget $(\omega_{g(j)}, \omega_\tau, \omega_{\lambda_j})$.
```

Algorithm 2 Blocked Gibbs Sampling for GRwHS(Detailed)

```
Require: Data $(\boldsymbol{X}, \boldsymbol{y})$, groups $\{\mathcal{G}_g\}_{g=1}^{G}$, hyperparameters $(c, \tau_0, \eta, s_0)$, iterations $T$.
Initialize $\boldsymbol{\beta}^{(0)}$, scales $\{\lambda_j^{(0)}, \phi_g^{(0)}, \tau^{(0)}, \sigma^{2(0)}\}$, and auxiliaries $\{\xi_j^{(0)}, \xi_\tau^{(0)}, \xi_\sigma^{(0)}\}$.
for $t = 1$ to $T$ do
    Update regression coefficients $\boldsymbol{\beta}$:
    Compute $d_{jj} \leftarrow (\phi_{g(j)}^2 \tau^2 \tilde{\lambda}_j^2 \sigma^2)^{-1}$ and form $\boldsymbol{D}_\beta \leftarrow \mathrm{diag}(d_{jj})$.
    Work with scaled precision $\boldsymbol{\Sigma}_\beta^{-1} = \sigma^{-2}\big(\boldsymbol{X}^\top \boldsymbol{X} + \mathrm{diag}((\phi^2\tau^2\tilde{\lambda}^2)^{-1})\big)$; compute $\boldsymbol{\mu}_\beta$.
    Sample $\boldsymbol{\beta}^{(t)} \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$ (using Woodbury if $p > n$).
    Update local scales $\boldsymbol{\lambda}$ for $j = 1, \dots, p$:
    Sample $u_j = \log \lambda_j$ via slice/MH from log-target proportional to:
        $-\log \tilde{\lambda}_j - \frac{\beta_j^2}{2\phi_{g(j)}^2 \tau^2 \tilde{\lambda}_j^2 \sigma^2} + \log \mathrm{C}^+(\lambda_j \mid 0, 1) + u_j$.
    Optionally, sample auxiliary $\xi_j \sim \mathrm{InvGamma}(1, 1 + 1/\lambda_j^2)$.
    Update group scales $\boldsymbol{\phi}$ for $g = 1, \dots, G$:
    Compute $p_g \leftarrow |\mathcal{G}_g|$, set $\eta_g \leftarrow \eta/\sqrt{p_g}$, and $S_g \leftarrow \frac{1}{\tau^2\sigma^2} \sum_{j \in \mathcal{G}_g} \frac{\beta_j^2}{\tilde{\lambda}_j^2}$.
    Sample $\theta_g \sim \mathrm{GIG}(\frac{1}{2} - \frac{p_g}{2}, S_g, \frac{1}{\eta_g^2})$ *using a RoU sampler robust for $\lambda \le 0$*; set $\phi_g \leftarrow \sqrt{\theta_g}$.
    Update global scale $\tau$:
    Sample $v = \log \tau$ via slice/MH from log-target:
        $-p \log \tau - \sum_j \log \tilde{\lambda}_j - \frac{1}{2\sigma^2} \sum_j \frac{\beta_j^2}{\phi_{g(j)}^2 \tau^2 \tilde{\lambda}_j^2} + \log \mathrm{C}^+(\tau \mid 0, \tau_0) + v$.
    Optionally, sample auxiliary $\xi_\tau \sim \mathrm{InvGamma}(1, 1/\tau_0^2 + 1/\tau^2)$.
    Update noise variance $\sigma^2$:
    Compute $RSS \leftarrow \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ and $S_\beta \leftarrow \sum_j \frac{\beta_j^2}{\phi_{g(j)}^2 \tau^2 \tilde{\lambda}_j^2}$.
    Set $B \leftarrow \frac{1}{2}(RSS + S_\beta) + \frac{1}{\xi_\sigma}$ and sample $\sigma^2 \sim \mathrm{InvGamma}(\frac{n+p+1}{2}, B)$.
    Optionally, sample auxiliary $\xi_\sigma \sim \mathrm{InvGamma}(1, 1/s_0^2 + 1/\sigma^2)$.
end for
Return posterior draws $\{\boldsymbol{\beta}^{(t)}, \dots\}_{t=1}^{T}$.
```

Algorithm 3 Structured Variational Inference (SVI) for GRwHS (Detailed)

```
Require: Variational family $q(\cdot)$; stepsize schedule $\{\rho_t\}$; minibatch size $m$; Hutchinson probes $m_{\mathrm{H}}$ (default 10–20); CG tolerance cg_tol ($10^{-3}$ default); MC samples for scale-priors $L$ (8–16).
Initialize $\vartheta^{(0)} = \{\boldsymbol{m}_g, \boldsymbol{S}_g, \boldsymbol{c}_g, \mu_{\log\phi_g}, v_{\log\phi_g}, \mu_{\log\lambda_j}, v_{\log\lambda_j}, a_j, \mu_{\log\tau}, v_{\log\tau}, \rho, \mu_{\log\sigma}, v_{\log\sigma}\}$.
for $t = 1$ to $T$ do
    Draw a random minibatch of rows, $\mathcal{I}_t \subset \{1, \dots, n\}$.
    Sample the shared factor $u \sim \mathcal{N}(0,1)$ and reparameterized samples from $q_{\vartheta^{(t)}}$.
    Construct the ELBO using analytic terms (minibatch likelihood; group block via Lemma 2.1) and MC/control-variates for Half-Cauchy log-priors ($L$ samples).
    *(Optional, diagnostics)* If shrinkage in coefficient-space is needed, estimate
      $A^{(t)}(\boldsymbol{s}) \leftarrow \big((\sigma^{(t)})^{-2}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{D}_\beta^{(t)}\big)\boldsymbol{s}$ with $\boldsymbol{D}_\beta^{(t)} = \mathrm{diag}(d_1^{(t)}, \dots, d_p^{(t)})$.
      For $r = 1{:}m_{\mathrm{H}}$: draw $\boldsymbol{r}^{(r)} \sim \{\pm 1\}^p$, solve $A^{(t)}(\boldsymbol{s}^{(r)}) = \boldsymbol{r}^{(r)}$ by CG at cg_tol,
      set $\boldsymbol{v}^{(r)} \leftarrow \boldsymbol{X}\boldsymbol{s}^{(r)}$, $\boldsymbol{u}^{(r)} \leftarrow (\sigma^{(t)})^{-2}\boldsymbol{X}^\top\boldsymbol{v}^{(r)}$,
      and $\widehat{\boldsymbol{\kappa}} \leftarrow \frac{1}{m_{\mathrm{H}}}\sum_r \boldsymbol{r}^{(r)} \odot \boldsymbol{u}^{(r)}$.
      (Use $p$-dim probes here; use $n$-dim probes only if estimating $\mathrm{diag}(H)$ in data-space.)
    Compute gradients $\widehat{\nabla}_{\mathrm{nat}}\mathcal{L}(\vartheta^{(t)})$;
      use CG for any linear solves involving $A^{(t)}$, and matrix-free matvecs for $X/X^\top$.
    Apply damping to covariance updates and early clipping to $(a_j, \rho)$.
    Update $\vartheta^{(t+1)} \leftarrow \vartheta^{(t)} + \rho_t\,\widehat{\nabla}_{\mathrm{nat}}\mathcal{L}(\vartheta^{(t)})$.
end for
Return optimized variational posterior $q_{\vartheta^{(T)}}$.
```

Algorithm 4 Post-processing for Shrinkage and Variance Diagnostics (Detailed)

```
Require: Posterior draws $\{\phi_g^{(t)}, \tau^{(t)}, \lambda_j^{(t)}, \sigma^{(t)}\}_{t=1}^T$, design matrix $\boldsymbol{X}$, hyperparameter $c$, small $\delta > 0$
    (e.g., $10^{-8}$), Hutchinson probes $m_{\mathrm{H}}$, CG tolerance cg_tol.
for each posterior draw $t = 1$ to $T$ do
    for each coefficient $j = 1$ to $p$ do
        $\tilde{\lambda}_j^{2(t)} \leftarrow \dfrac{c^2(\lambda_j^{(t)})^2}{c^2 + (\tau^{(t)})^2(\lambda_j^{(t)})^2}$            ▷ Regularized local scale
        $d_j^{(t)} \leftarrow \Big((\phi_{g(j)}^{(t)})^2(\tau^{(t)})^2\tilde{\lambda}_j^{2(t)}(\sigma^{(t)})^2\Big)^{-1}$            ▷ Prior precision
    end for
    Coefficient-space shrinkage via Hutchinson+CG:
   $\boldsymbol{D}_\beta \leftarrow \mathrm{diag}\big(d_1^{(t)}, \ldots, d_p^{(t)}\big)$
   Define $A^{(t)}(\boldsymbol{s}) = \big((\sigma^{(t)})^{-2}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{D}_\beta\big)\,\boldsymbol{s}$
   for $r = 1, \ldots, m_{\mathrm{H}}$ do
      Draw Rademacher probe $\boldsymbol{r}^{(r)} \sim \{\pm 1\}^p$
      Solve $A^{(t)}(\boldsymbol{s}^{(r)}) = \boldsymbol{r}^{(r)}$ by CG at tolerance cg_tol
      $\boldsymbol{v}^{(r)} \leftarrow \boldsymbol{X}\boldsymbol{s}^{(r)};\ \boldsymbol{u}^{(r)} \leftarrow (\sigma^{(t)})^{-2}\boldsymbol{X}^\top\boldsymbol{v}^{(r)}$
   end for
   $\widehat{\boldsymbol{\kappa}} \leftarrow \frac{1}{m_{\mathrm{H}}}\sum_{r=1}^{m_{\mathrm{H}}} \boldsymbol{r}^{(r)} \odot \boldsymbol{u}^{(r)}$
    for each coefficient $j = 1$ to $p$ do
        Correlation-aware shrinkage: $\kappa_j^{(t)} \leftarrow \widehat{\boldsymbol{\kappa}}_j$
        Normalized variance budget (priors-only):
   $a_{g(j)} \leftarrow 2\log\phi_{g(j)}^{(t)},\ a_\tau \leftarrow 2\log\tau^{(t)},\ a_{\lambda_j} \leftarrow 2\log\tilde{\lambda}_j^{(t)}$
   $a_k^\delta \leftarrow \mathrm{sign}(a_k)\cdot\max(|a_k|, \delta)$ for $k \in \{g(j), \tau, \lambda_j\}$
   $\omega_{g(j)}^{(t)} \leftarrow \dfrac{a_{g(j)}^\delta}{a_{g(j)}^\delta + a_\tau^\delta + a_{\lambda_j}^\delta},\ \omega_\tau^{(t)} \leftarrow \dfrac{a_\tau^\delta}{a_{g(j)}^\delta + a_\tau^\delta + a_{\lambda_j}^\delta},\ \omega_{\lambda_j}^{(t)} \leftarrow \dfrac{a_{\lambda_j}^\delta}{a_{g(j)}^\delta + a_\tau^\delta + a_{\lambda_j}^\delta}$
        $r_j^{(t)} \leftarrow \dfrac{(\tau^{(t)})^2(\lambda_j^{(t)})^2}{c^2}$            ▷ Slab vs. spike indicator
    end for
    for each group $g = 1$ to $G$ do
        $\mathrm{edf}_g^{(t)} \leftarrow \sum_{j\in\mathcal{G}_g}\kappa_j^{(t)}$            ▷ Group effective degrees of freedom
    end for
end for
Return posterior summaries of $\{\kappa_j, \omega_{g(j)}, \omega_\tau, \omega_{\lambda_j}, r_j\}_{j=1}^p$ and $\{\mathrm{edf}_g\}_{g=1}^G$, including medians,
    credible intervals, and $\Pr(r_j > 1)$.
```

Appendix A: Proof of Lemma 2.1

Let (β, W) be jointly Gaussian with mean (m_β, μ_W), variances (v_β, s_W^2), and covariance $\Sigma_{\beta W}$. Write $T = -W$, which is Gaussian with mean $-\mu_W$ and variance s_W^2. By the mgf of a joint normal, for any jointly Gaussian (X, Y),

$$\mathbb{E}\big[X^2 e^{tY}\big] = \exp\Big(t\mu_Y + \tfrac{1}{2}t^2 s_Y^2\Big)\Big(v_X + (m_X + t\,\Sigma_{XY})^2\Big).$$

Taking $(X, Y, t) = (\beta, T, 1)$ gives

$$\mathbb{E}[\beta^2 e^T] = \exp\Big(- \mu_W + \tfrac{1}{2} s_W^2 \Big) \Big(v_\beta + (m_\beta - \Sigma_{\beta W})^2 \Big),$$

and hence $\mathbb{E}[\beta^2 e^{-W}]$ equals the same right-hand side, proving the claim. $\square$