Hierarchical Bayesian Estimation of Covariance Matrices

Daniel Yekutieli

Tel Aviv University

ISBA 2026, Nagoya, Japan

Joint work with Daniel Xiang, Malgorzata Bogdan, Jonas Wallin:

- Equivariance under the General Linear Group of $p \times p$ nonsingular matrices ($GL(p)$) of estimating covariance matrices from Normal data is well known (Stein, 1961; Berger, 2005)
- However, $GL(p)$-equivariant estimators carry considerably larger risks than shrinkage estimators (Stein, 1976; Haff, 1980, 1991; Ledoit and Wolf, 2004, 2020; Donoho, Gavish, and Johnstone, 2018) that are equivariant under the subgroup of orthogonal matrices ($\mathrm{O}(p)$).
- For known eigenvalue vector (oracle), we derive Bayes rules with respect to Haar measure on $\mathrm{O}(p)$ for estimating covariance and precision matrices for invariant loss functions
- We evaluate oracle Bayes rules by a Gibbs sampling algorithm – they serve as theoretical risk minimizing benchmarks
- We introduce a hierarchical Bayes model that places a finite Pólya tree prior on the eigenvalue distribution and use Gibbs sampling to generate posterior draws, yielding both shrinkage estimates for the eigenvalues and approximations to the oracle Bayes rules.

Notations

- Observe $\mathbf{X}_{n\times p}$ with i.i.d. rows $\mathbf{X}^i \sim N(\mathbf{0}, \mathbf{\Sigma})$.
- The eigendecomposition is central:

$$\boldsymbol{\lambda} = (\lambda_1 \geq \cdots \geq \lambda_p > 0),\ \mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^\top,\ \mathbf{\Omega} = \mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{\Gamma}^\top,$$

  where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ and $\mathbf{\Gamma} \in \mathrm{O}(p)$.

- The sample covariance matrix $\mathbf{S} = \mathbf{X}^\top\mathbf{X}/n$ with eigedecomposition $\mathbf{S} = \mathbf{V}\mathbf{L}\mathbf{V}^\top$, $\mathbf{l} = (l_1 \geq \cdots \geq l_p)$.
- We consider three equivariant matrix loss functions:

L_0	Frobenius	$\mathrm{tr}\left((\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})^2\right)$
L_1	Stein loss	$\mathrm{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}) - \log\det(\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}) - p$
L_2	Squared Stein	$\mathrm{tr}\left((\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1} - I)^2\right)$

Equivariance Under $\mathrm{O}(p)$

For $\mathbf{R} \in \mathrm{O}(p)$, the transformation $g_{\mathbf{R}}$ acts on data, $g_{\mathbf{R}}(\mathbf{X}) = \mathbf{X}\mathbf{R}$, $\bar{g}_{\mathbf{R}}$ acts on parameters: $\bar{g}_{\mathbf{R}}(\boldsymbol{\Sigma}) = \mathbf{R}^\top \boldsymbol{\Sigma} \mathbf{R}$ and $\bar{g}_{\mathbf{R}}(\boldsymbol{\Omega}) = (\bar{g}_{\mathbf{R}}(\boldsymbol{\Sigma}))^{-1}$.

Model equivariance: $\mathbf{X}^i \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \;\Rightarrow\; \mathbf{X}^i\mathbf{R} \sim N(\mathbf{0}, \mathbf{R}^\top \boldsymbol{\Sigma} \mathbf{R})$.

Estimator equivariance: $\hat{\boldsymbol{\Sigma}}(g_{\mathbf{R}}(\mathbf{X})) = \bar{g}_{\mathbf{R}}(\hat{\boldsymbol{\Sigma}}(\mathbf{X}))$.

Loss invariance: $\hat{\boldsymbol{\Sigma}}(g_{\mathbf{R}}(\mathbf{X})) = \bar{g}_{\mathbf{R}}(\hat{\boldsymbol{\Sigma}}(\mathbf{X}))$.

Transitivity: Writing $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top = \bar{g}_{\boldsymbol{\Gamma}^\top}(\boldsymbol{\Lambda})$, shows that $\mathrm{O}(p)$ acts transitively on $\mathcal{M}_{\boldsymbol{\lambda}} = \{\boldsymbol{\Sigma} : \text{eigenvalues} = \boldsymbol{\lambda}\}$.

Theorem

*If the model, estimator, and loss are equivariant/invariant under $\mathrm{O}(p)$, and $\mathrm{O}(p)$ is transitive on $\mathcal{M}_{\boldsymbol{\lambda}}$ then risk of the estimator is **constant**.*

Structure of Orthogonally Equivariant Estimators

Proposition

If estimator is equivariant under* $\mathrm{O}(p)$*, then*

$$\hat{\mathbf{\Sigma}}(\mathbf{S}) = \mathbf{V}\,\hat{\mathbf{\Sigma}}(\mathbf{L})\,\mathbf{V}^\top, \quad \hat{\mathbf{\Sigma}}(\mathbf{L}) = \mathrm{diag}(d_1, \ldots, d_p).$$

Moreover, if $n \leq p-2$*, then* $d_{n+1} = \cdots = d_p$*.*

Examples of orthogonally equivariant estimators:

- MLE*: $\hat{\lambda}_j = l_j$
- Stein shrinkage: $\hat{\lambda}_j = l_j/(n+p+1-2j)$
- Haff (1980) empirical Bayes: $\hat{\lambda}_j \propto l_j +$ (pooled correction)
- Ledoit & Wolf (2004), (2020): linear and nonlinear shrinkage
- Donoho, Gavish and Johnstone (2018) estimator with non-linear shrinkage function depending on the loss function and the limiting aspect ratio $n/p \to \gamma \in (0, 1]$. *: hard thresholding at optimal threshold

Oracle Generative Model – assume fixed known $\boldsymbol{\lambda}$

$$\boldsymbol{\Gamma} \sim \pi^{\text{Haar}} \text{ on } \mathrm{O}(p), \quad \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top, \quad \mathbf{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})^{\otimes n}.$$

- posterior distribution of $\boldsymbol{\Gamma}$ is a *matrix Bingham distribution*:

$$\pi^{\text{Haar}}(\boldsymbol{\Gamma} \mid \mathbf{x}, \boldsymbol{\lambda}) \;\propto\; \exp\Big(\mathrm{tr}\Big(-\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}^\top (n\mathbf{S})\boldsymbol{\Gamma}/2\Big)\Big).$$

- The sample eigenvector matrix $\mathbf{V}$ is a **mode** of the posterior.
- The posterior is **equivariant** under $\mathrm{O}(p)$.
- The oracle Bayes rule is the estimator minimizing the average risk, given by

$$\hat{\Sigma}^{\text{Haar}}(\mathbf{x}, \boldsymbol{\lambda}) := \arg\min_{\hat{\boldsymbol{\Sigma}}} \Big[E_{\boldsymbol{\Gamma} \sim \pi^{\text{Haar}}(\boldsymbol{\Gamma} \mid \mathbf{x}, \boldsymbol{\lambda})} L(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})\Big].$$

Theorem

Under any $\mathrm{O}(p)$-invariant loss, the oracle Bayes rule is the ***equivariant estimator with the smallest risk****.*

Oracle Bayes Rules:

The oracle Bayes rule is $\hat{\mathbf{\Sigma}}(\mathbf{L}) = \mathrm{diag}(d_1, \ldots, d_p)$.

Frobenius loss $(n \geq p)$:

$$d_k = \sum_{j=1}^{p} \lambda_j \, \mathbb{E}[\Gamma_{kj}^2 \mid \mathbf{S} = \mathbf{L}]$$

Stein loss $(n \geq p)$:

$$d_k = \left(\sum_{j=1}^{p} \lambda_j^{-1} \, \mathbb{E}[\Gamma_{kj}^2 \mid \mathbf{S} = \mathbf{L}] \right)^{-1}$$

Squared Stein too complicated....

Notice that...

Posterior mean row entries $(\Gamma_{k1}^2, \cdots, \Gamma_{kp}^2)$ sum to 1, as
$\mathbf{S} = \mathbf{L} \Rightarrow \mathbf{V} = \mathbf{I}$ then the posterior mode of $\mathbf{\Gamma}$ is $\mathbf{I}$, where for large n
$(\Gamma_{k1}, \cdots, \Gamma_{kp})$ converges to the k-th unit vector

More Oracle Bayes Rules:

Oracle Bayes rule for estimating the precision matrix is

$$\hat{\boldsymbol{\Omega}}^{\text{Haar}}(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{V}\,\text{diag}(\hat{\omega}_1, \ldots, \hat{\omega}_p)\mathbf{V}^\top$$

where $\hat{\omega}_k = d_k(\lambda_1^{-1}, \cdots, \lambda_p^{-1})$ in expression for $\hat{\boldsymbol{\Sigma}}(\mathbf{L})$.

and for the no-data case ($n = 0$)

Bayes rules collapse to a fixed constant diagonal:

Loss	$\hat{\Sigma}$ **diagonal**	$\hat{\Omega}$ **diagonal**
Frobenius	$\frac{\sum \lambda_j}{p}$	$\frac{\sum \lambda_j^{-1}}{p}$
Stein	$\frac{p}{\sum \lambda_j^{-1}}$	$\frac{p}{\sum \lambda_j}$
Squared Stein	$\frac{\sum \lambda_j^{-1}}{\sum \lambda_j^{-2}}$	$\frac{\sum \lambda_j}{\sum \lambda_j^2}$

Sampling the Oracle Posterior

We use Gibbs sampling based on the Diaconis & Shahshahani (1987) subgroup algorithm to produce posterior samples from

$$\pi^{\mathrm{Haar}}(\mathbf{\Gamma} \mid \mathbf{x}, \boldsymbol{\lambda})$$

The next two slides illustrate:

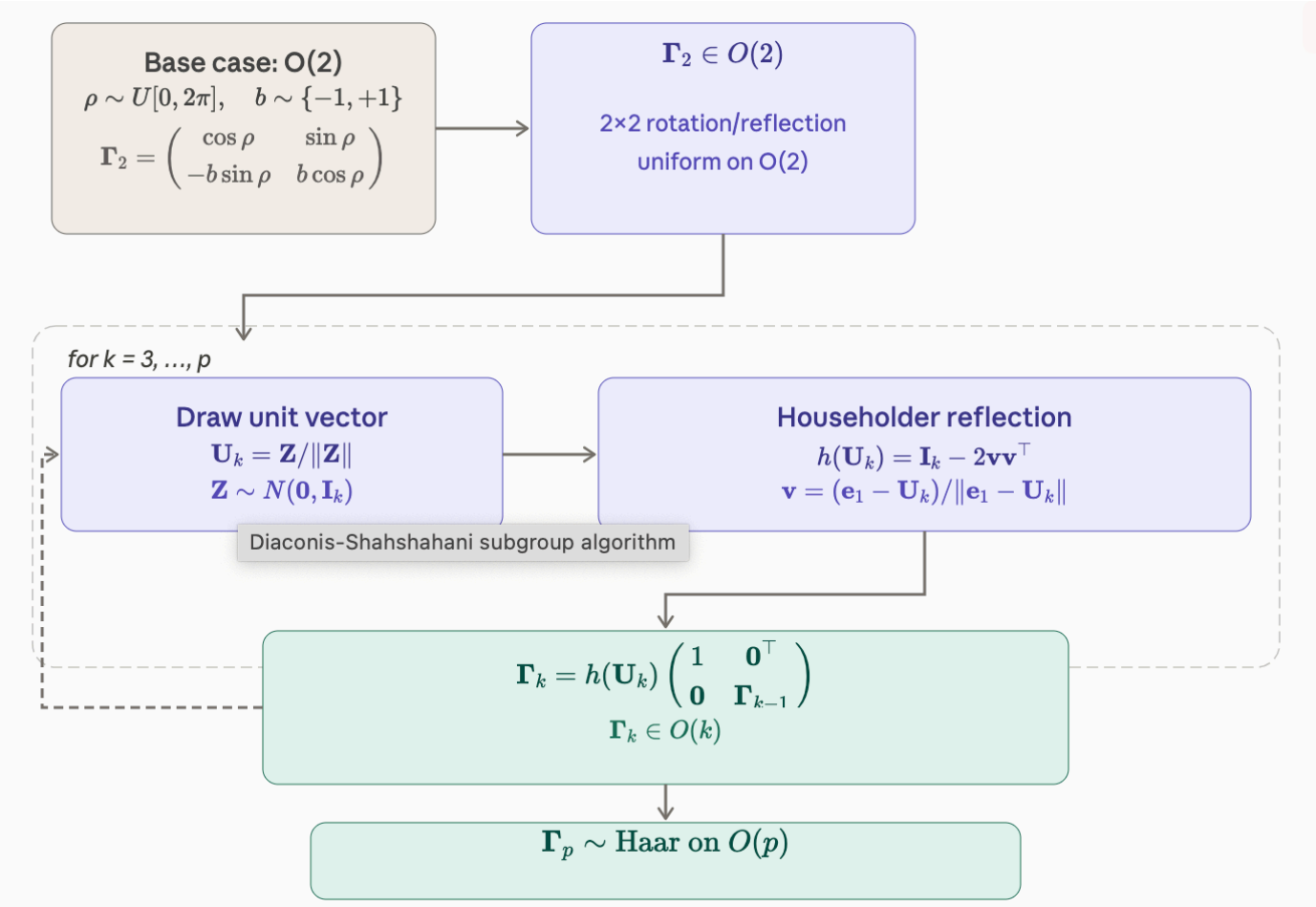
1. Subgroup algorithm

Samples $\mathrm{O}(p)$ via successive Householder reflections.

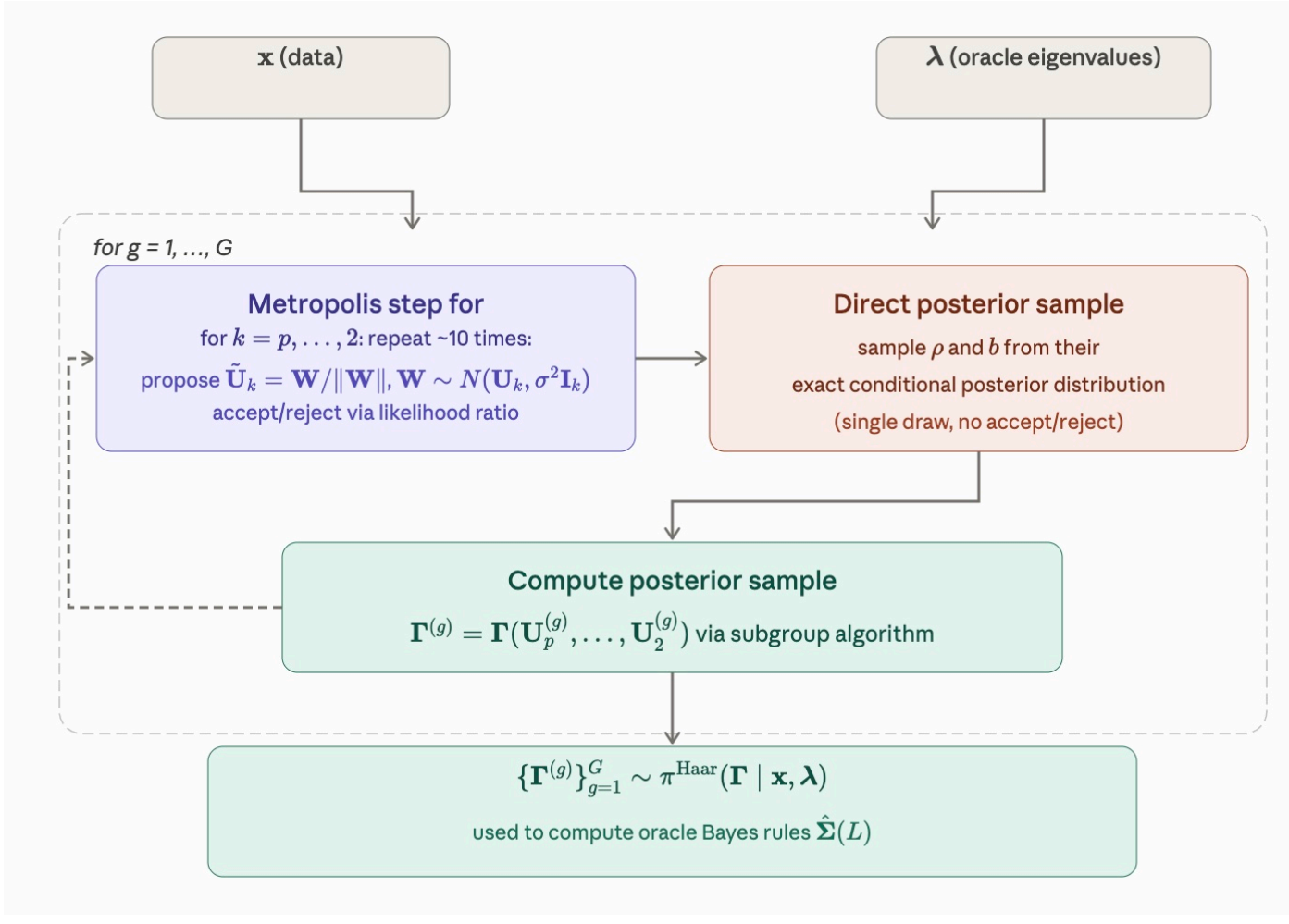
2. Oracle Gibbs sampler

yield posterior draws $\mathbf{\Gamma}^{(1)}, \ldots, \mathbf{\Gamma}^{(G)}$ from $\pi^{\mathrm{Haar}}(\mathbf{\Gamma} \mid \mathbf{x}, \boldsymbol{\lambda})$.

Diaconis–Shahshahani Subgroup Algorithm:



Oracle Gibbs Sampler



From Oracle to hBayes

In practice, $\boldsymbol{\lambda}$ is unknown. We place a L level Finite Pólya Tree (FPT) prior on the eigenvalues yielding generative model:

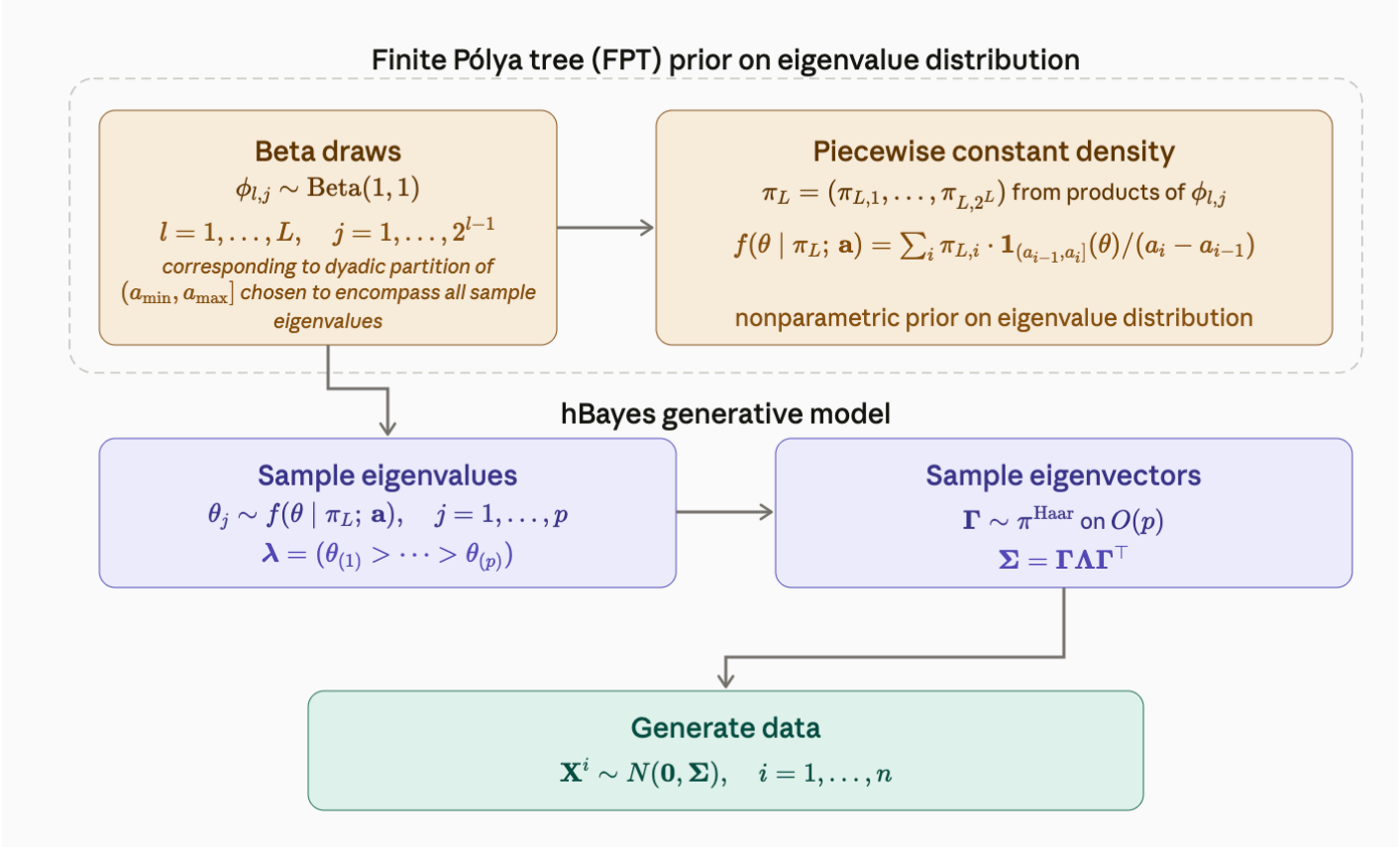
$$\boldsymbol{\lambda} \sim \text{FPT} \longrightarrow \boldsymbol{\Gamma} \sim \pi^{\text{Haar}} \longrightarrow \mathbf{X} \sim N(\mathbf{0}, \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top)^{\otimes n}$$

Why FPT?

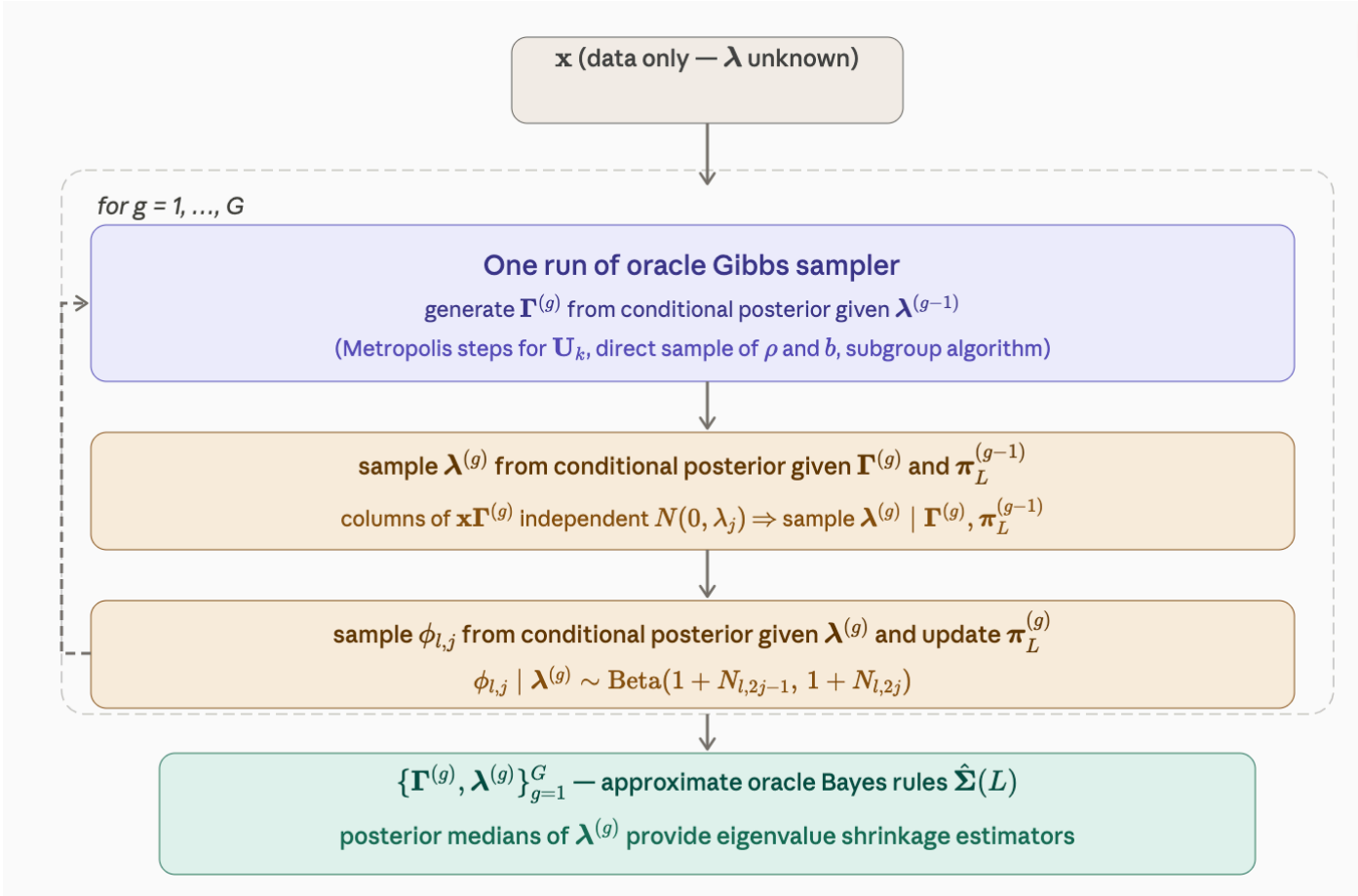
- Nonparametric — 2^L parameter step-function density on a dyadic partition of $(a_{\min}, a_{\max}]$ chosen to cover all sample eigenvalue.
- Closed-form conjugate posterior updates for the split probabilities $\phi_{l,j}$.
- Strong regularisation — stable eigenvalue distribution estimates even for small n.

The next slides illustrate the hBayes generative model and the hBayes Gibbs sampler we use to sample it...

hBayes generative model



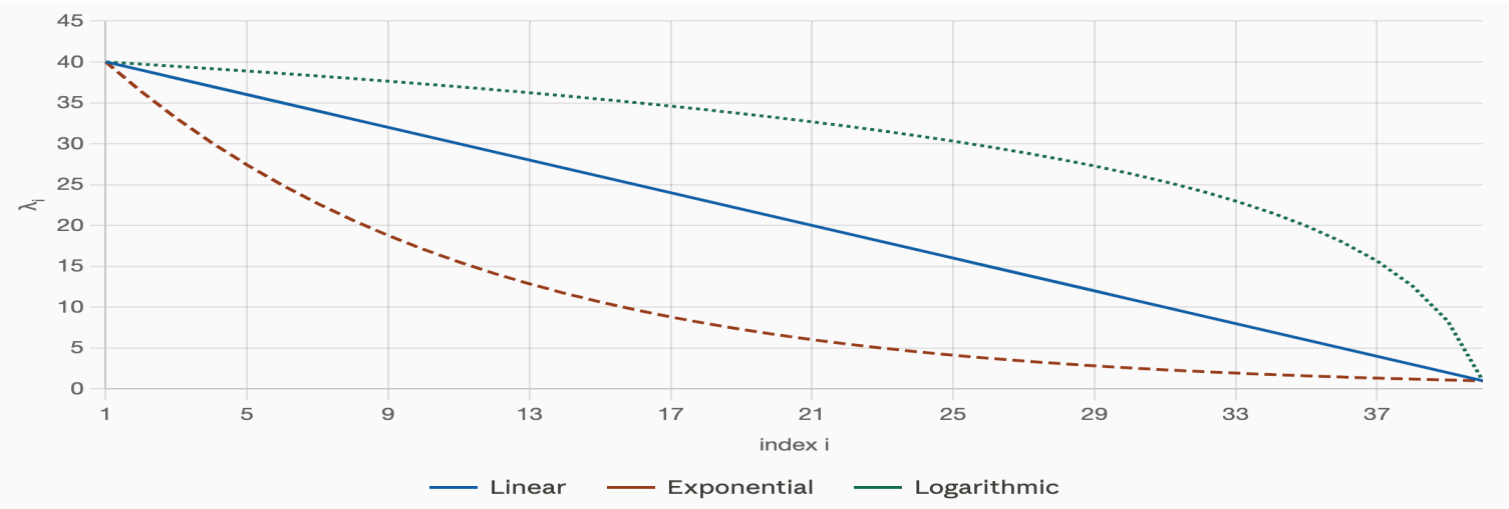
hBayes Gibbs sampler



Simulation Setup

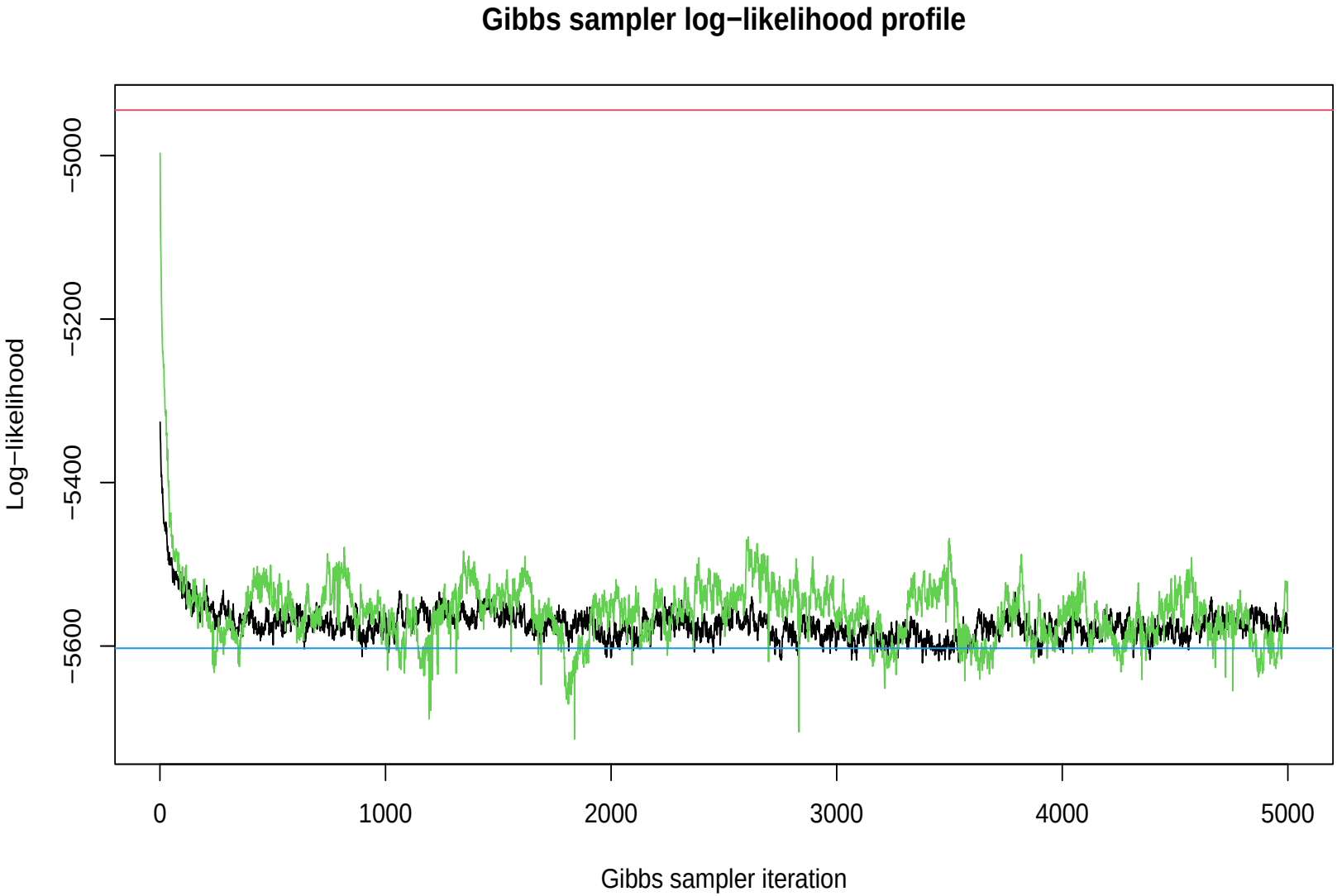
Sample sizes: $n = 50$ and $n = 100$ **while** $p = 40$.
Three eigenvalue configurations ($\lambda_1 = 40, \cdots, \lambda_{40} = 1$):

- **Linear:** $\lambda_i = (p+1) - i$
- **Exponential:** $\lambda_i = a \cdot e^{-bi}$
- **Logarithmic:** $\lambda_i = a \cdot \log(p + 1 - i) + 1$



Number of FPT levels: $L = 6$ therefore model $\boldsymbol{\lambda}$ distribution with

$$f(\theta|\boldsymbol{\pi}_6; \mathbf{a}) = \pi_{6,1} \cdot \frac{I_{(a_0, a_1]}(\theta)}{a_1 - a_0} + \pi_{6,2} \cdot \frac{I_{(a_1, a_2]}(\theta)}{a_2 - a_1} + \cdots + \pi_{6,64} \cdot \frac{I_{(a_63, a_{64}]}(\theta)}{a_{64} - a_{63}}$$



Eigenvalue configurations, $p = 40$ $\log(f(\mathbf{x} \mid \cdot))$ for simulated $\mathbf{x}$ with $p = 40$ and $n = 50$, linear $\boldsymbol{\lambda} = (40, 39, \ldots, 1)$. The red line is $\log(f(\mathbf{x} \mid \mathbf{S}))$, blue line is $\log(f(\mathbf{x} \mid \boldsymbol{\Sigma}))$. The black and green curves $\log(f(\mathbf{x} \mid \boldsymbol{\Sigma}^{(g)}))$ for iterations $g = 1, \ldots, 5000$ of the oracle and hBayes samplers.

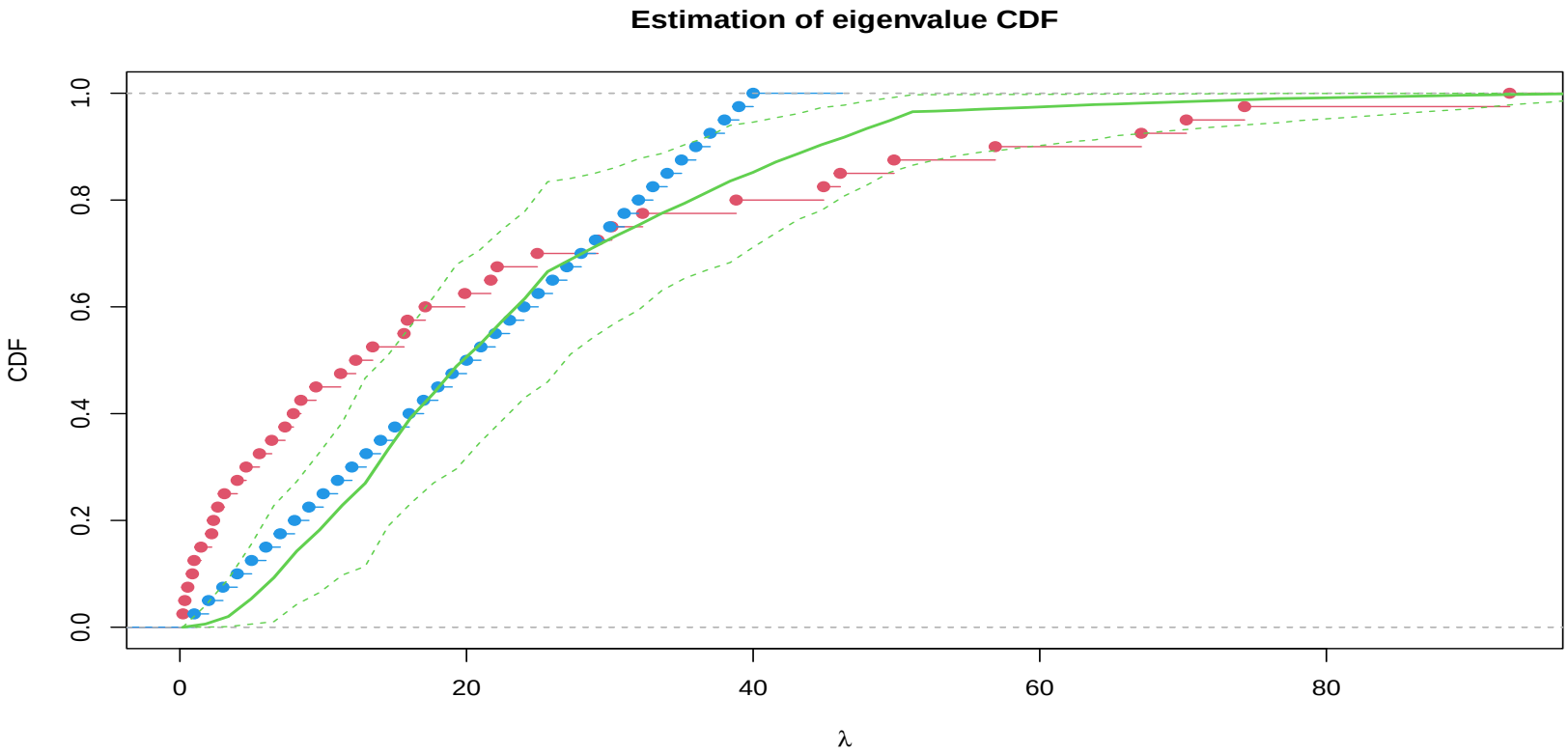


Figure: Same simulation. Blue curve is eCDF of $\lambda_1, \ldots, \lambda_p$. Red curves is eCDF of $l_1, \ldots, l_p$. Green curves are hBayes $\boldsymbol{\lambda}$ CDF estimates: posterior median (solid line) and 0.05 and 0.95 quantiles of $\pi_{6,1}^{(g)} + \cdots + \pi_{6,j}^{(g)}$ with $j = 1, \ldots, 64$, for $g = 1501, \ldots, 5000$.

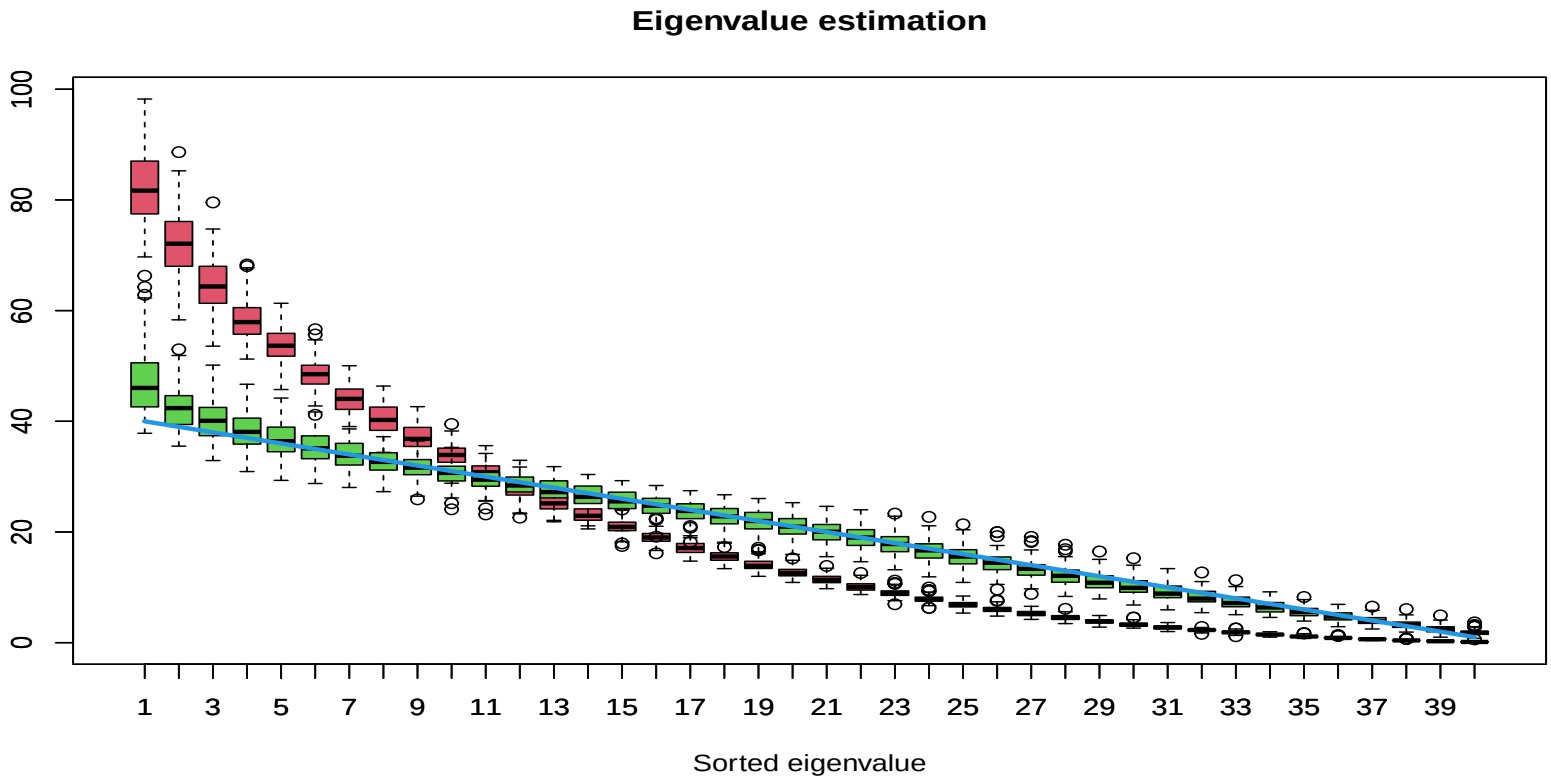


Figure: Boxplots of distribution of the sorted eigenvalue estimates across 100 simulated datasets with p = 40 and n = 50 and linear $\boldsymbol{\lambda} = (40, 39, \ldots, 1)$. The red boxplots display the distribution of the sample eigenvalues. The green boxplots display the distribution of the hBayes estimates: medians of the sorted eigenvalues from the hBayes sampler. The underlying sorted eigenvalues are shown in blue.

Risk Comparison: Covariance Matrix Estimation

Two hBayes approximations

hB-1 Posterior medians of $\boldsymbol{\lambda}$; use hBayes draws $\boldsymbol{\Gamma}^{(g)}$

hB-2 Estimate $\hat{\boldsymbol{\lambda}}$ from hBayes, re-run oracle sampler at $\hat{\boldsymbol{\lambda}}$ for $\boldsymbol{\Gamma}^{(g)}$

Simulation with $p = 40$, $n = 80$, linearly decreasing $\boldsymbol{\lambda}$, 100 datasets:

Stein loss L_1

Estimator	Mean	SE
Oracle	1.76	0.01
hBayes-1	1.86	0.02
hBayes-2	1.87	0.02
Haff	2.24	0.03
Stein iso	2.28	0.03
MLE	12.66	0.06

Squared Stein loss L_2

Estimator	Mean	SE
Oracle	3.19	0.01
hBayes-1	3.49	0.04
hBayes-2	3.51	0.04
Haff	4.10	0.04
Stein iso	4.59	0.11
MLE	20.46	0.15

Risk Comparison: Precision Matrix Estimation

Simulation with $p = 40$, $n = 80$, linearly decreasing $\boldsymbol{\lambda}$, 100 datasets:

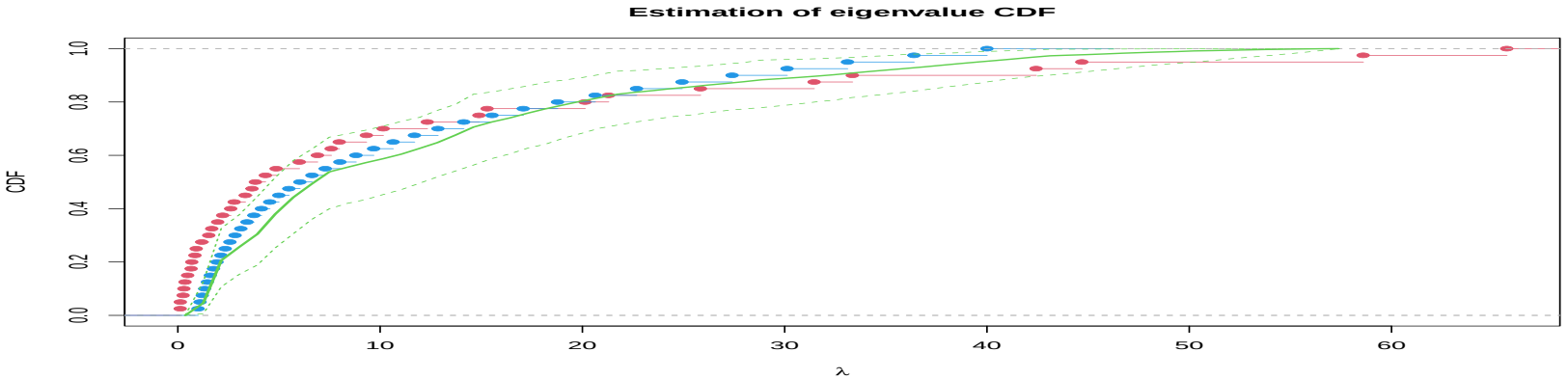
Stein loss L_1

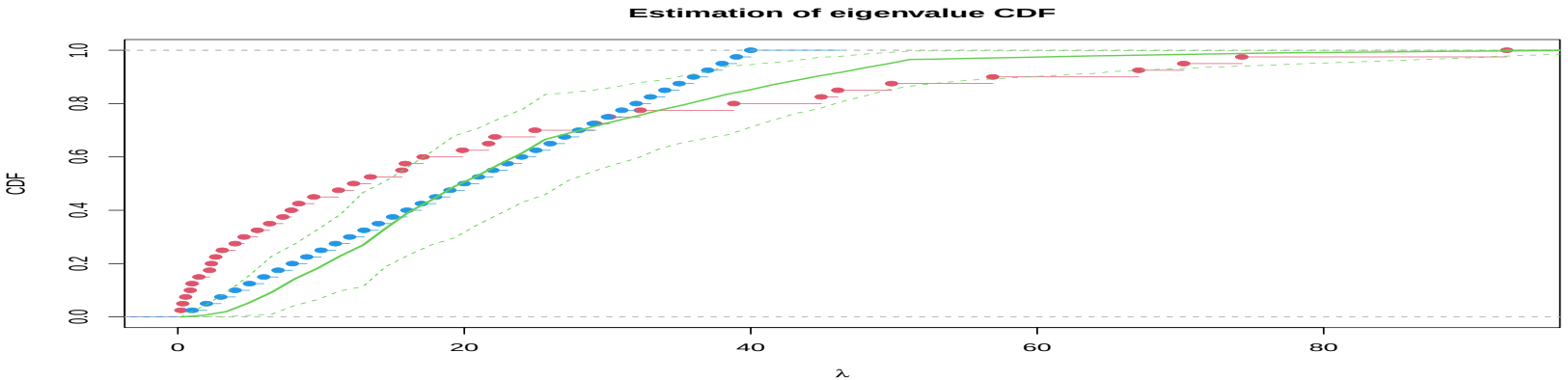
Estimator	Mean	SE
Oracle	1.77	0.01
hBayes-1	1.84	0.01
hBayes-2	1.85	0.01
glasso	2.71	0.04
MLE	29.50	0.29

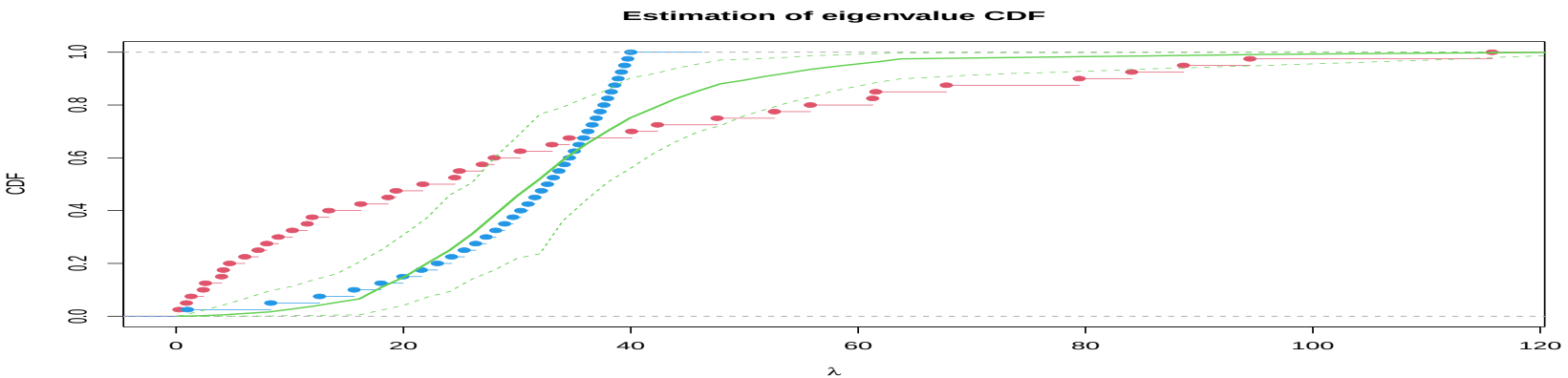
Squared Stein loss L_2

Estimator	Mean	SE
Oracle	3.26	0.01
hBayes-1	3.36	0.01
hBayes-2	3.38	0.01
glasso	6.12	0.14
MLE	224.38	4.36

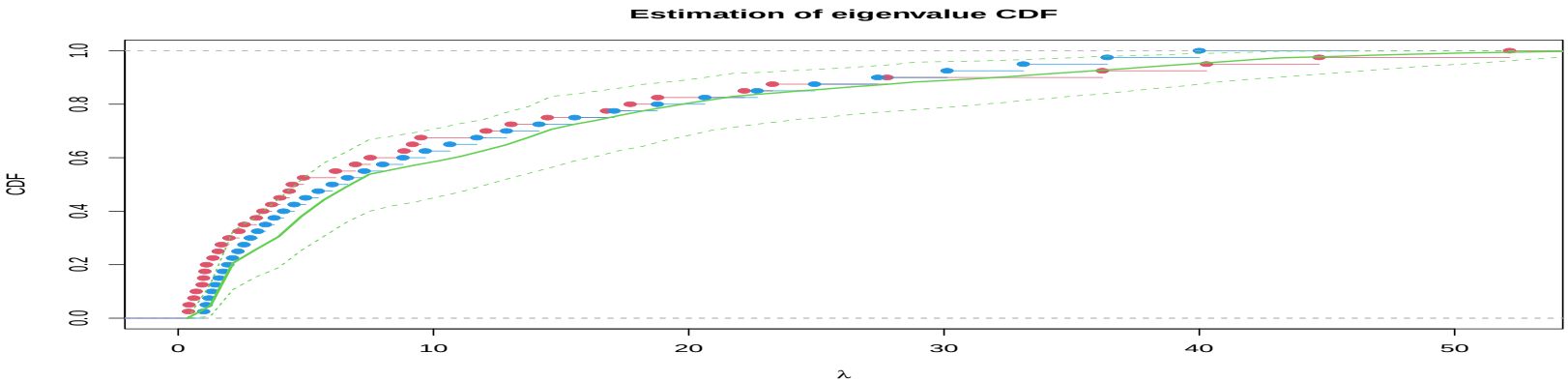
Exponential / Linear / Logarithmic λ for $n = 50$

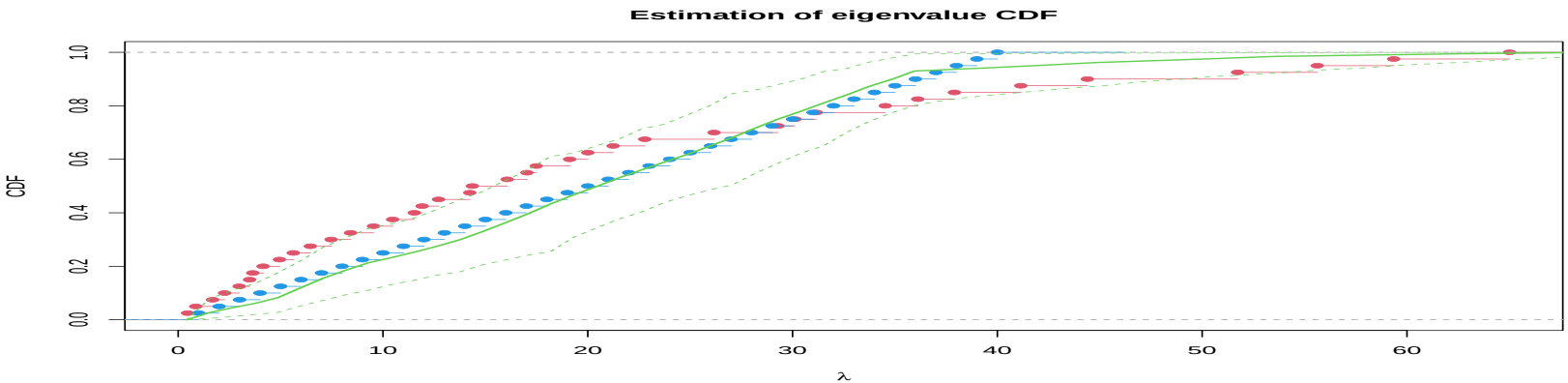


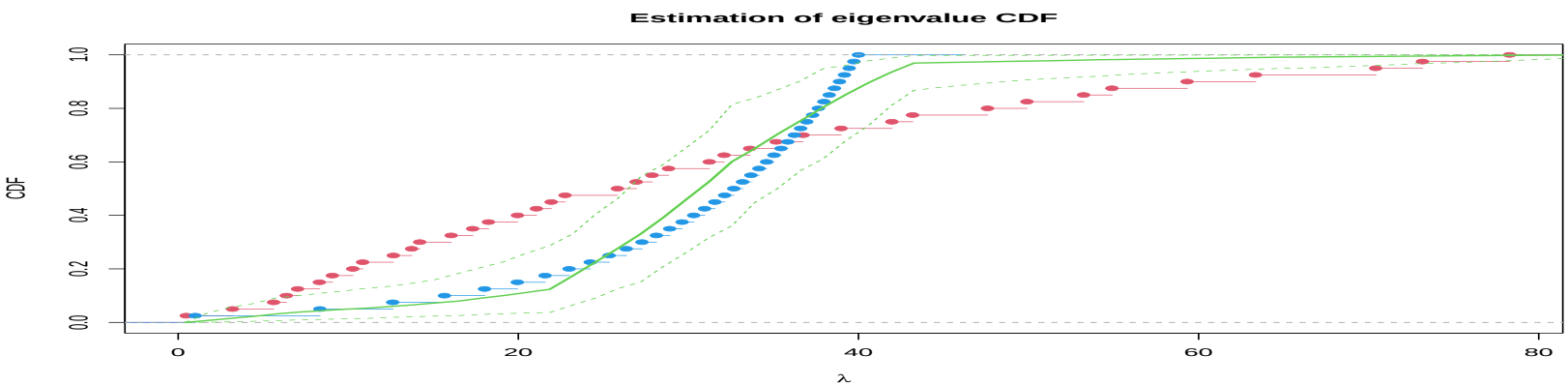




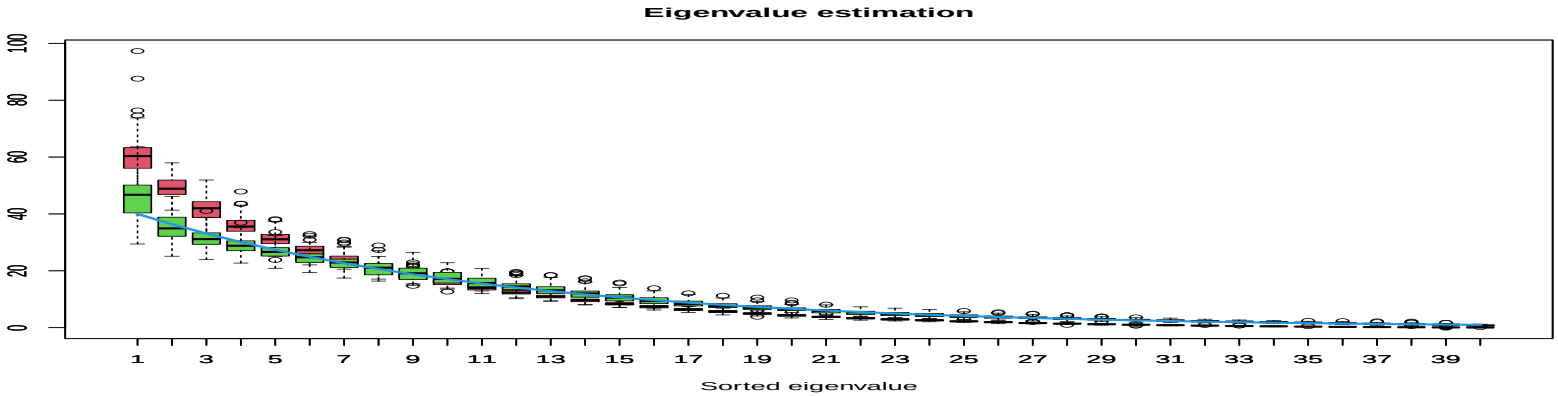
Exponential / Linear / Logarithmic λ for $n = 100$

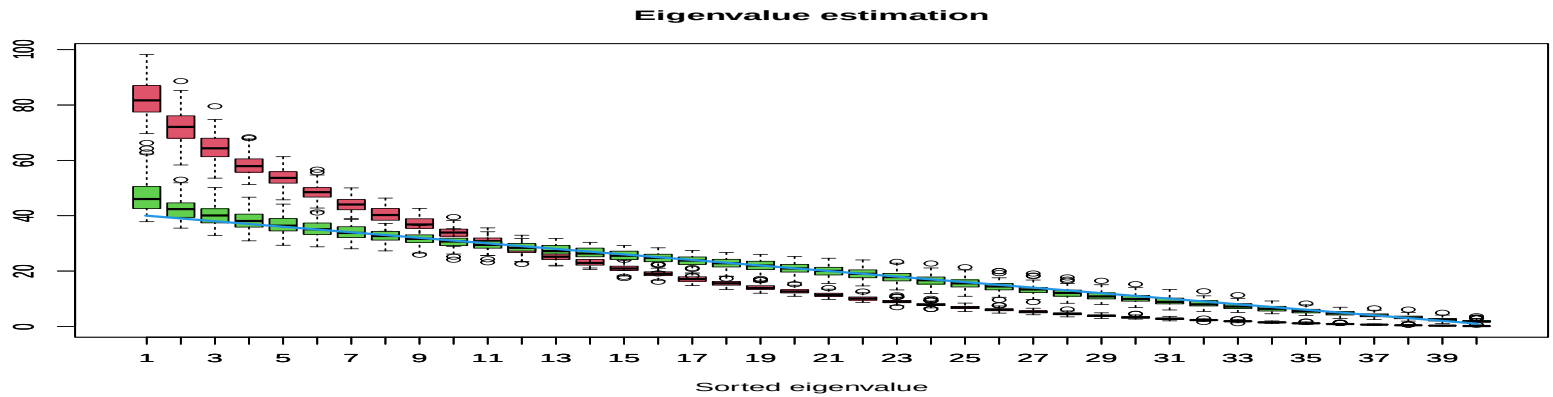


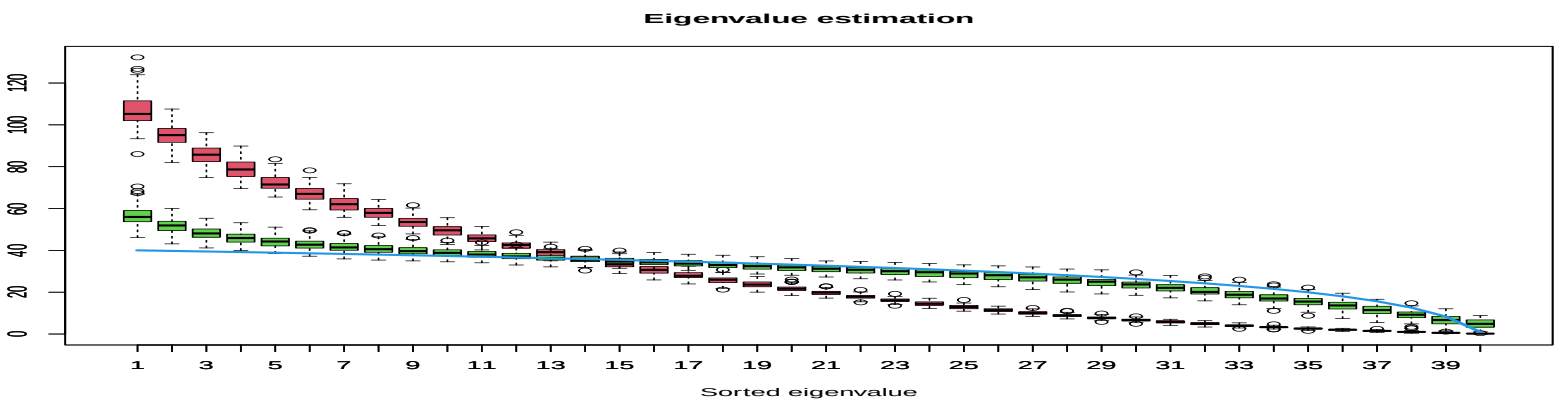




Exponential / Linear / Logarithmic λ for $n = 50$







Discussion

- Knowing the sorted eigenvalue vector $\boldsymbol{\lambda}$ (the oracle) without specifying $\boldsymbol{\Gamma}$ is equivalent to knowing the discrete distribution of the eigenvalues.
- The hBayes model with continuous FPT prior on the eigenvalues recovers the **general form of the eigenvalue distribution**, approximates the oracle Bayes rules and yields eigenvalue shrinkage estimators.
- **Implementation recommendation:** when approaching a new problem, run a small simulation to evaluate the risk gap between common shrinkage estimators and the oracle Bayes rule. If the gap is substantial, we recommend implementing the hBayes approach.
- Preprint available at `arxiv.org/abs/2606.24751`.

Thank you!