Resampling PCA & GP Inference

Manfred Opper

(ISIS, University of Southampton)

Motivation

- Construct "simple" intractable GP model
- Study approximate (EC/EP) inference
- "MC" conceptually simple
- Get a quantitative idea why EC inference works.

Resampling (Bootstrap)

Estimate average case properties (test errors) of statistical estimators based on a single dataset

$D_0 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3\}$

Bootstrap: Resample with replacement $\rightarrow$ Generate pseudo data.

$D_1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_2\}$, $D_2 = \{\mathbf{y}_1, \mathbf{y}_1, \mathbf{y}_1\}$, $D_3 = \{\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_3\}$, $\ldots$ etc

Problem: Each sample requires retraining of some learning algorithm.

Mapping to probabilistic model & Approximate inference: Only single training (inference) for single (effective) model required (Malzahn & Opper 2003).

PCA

- Goal: Project (d dimensional) data vectors $\mathbf{y} \to P_q[\mathbf{y}]$ on $q < d$ dimensional subspace with minimal reconstruction error $E||\mathbf{y} - P_q[\mathbf{y}]||^2$.

- Method: Approximate expectation by N training data D_0 given by the $(d \times N)$ matrix $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$. $\mathbf{y}_i \in R^d$.
  $d = \infty$ allowed (feature vectors).

  Optimal subspace spanned by eigenvectors $\mathbf{u}_l$ of *data covariance matrix*

  $$\mathbf{C} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^T$$

  corresponding to the q largest eigenvalues $\lambda_l \geq \lambda$.

Reconstruction Error

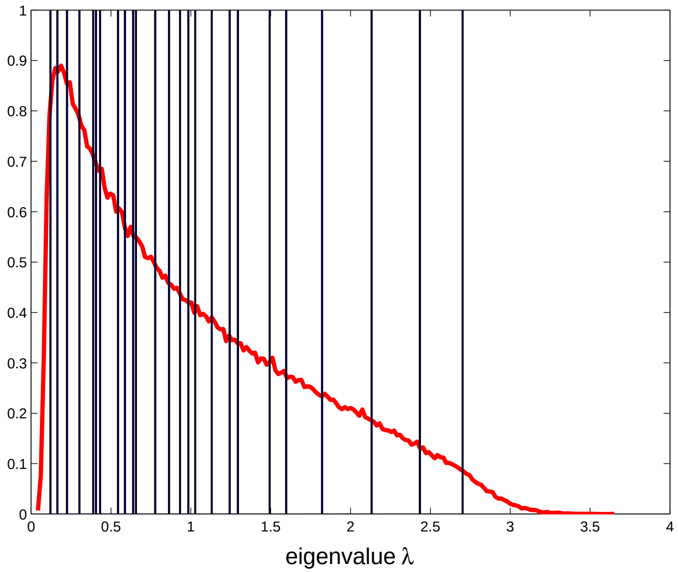
Expected reconstruction error (on novel data)

$$\varepsilon(\lambda) = \sum_{l:\lambda_l<\lambda} E(\mathbf{y}\cdot\mathbf{u}_l)^2$$

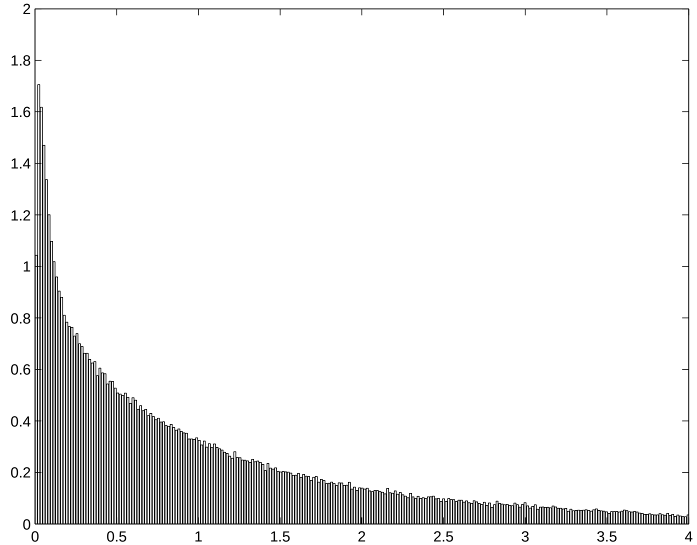
Resample averaged reconstruction error

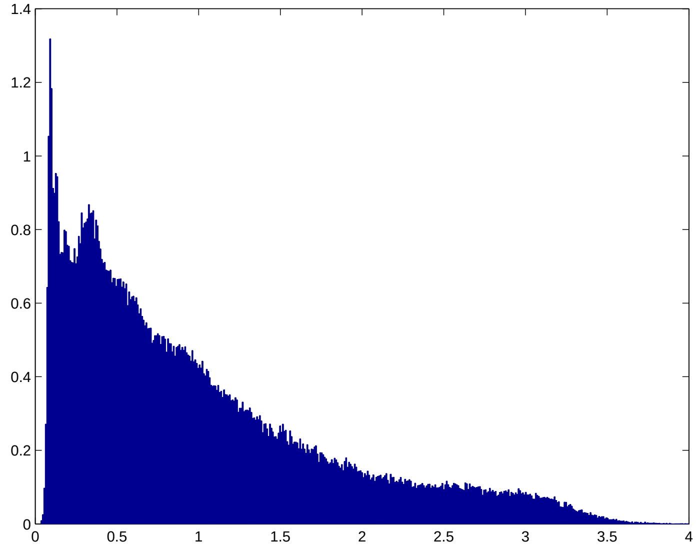
$$\mathcal{E}_r = \frac{1}{N_0}\mathbf{E}_{\mathbf{D}}\left[\sum_{\mathbf{y}_i\notin D;\lambda_l<\lambda} \mathrm{Tr}\left(\mathbf{y}_i\mathbf{y}_i^T\mathbf{u}_l\mathbf{u}_l^T\right)\right]$$

Bootstrap of density of Eigenvalues



Bootstrap ($N = 50$ random data, Dim $= 25$) $1\times$ and $3\times$ oversampled





The model

- Let $s_i = \#$ times $\mathbf{y}_i \in D$

- Diagonal *random matrix*

$$\mathbf{D}_{ii} = D_i = \frac{1}{\mu\Gamma}(s_i + \epsilon\delta_{s_i,0}) \qquad \mathbf{C}(\epsilon) = \frac{\Gamma}{N}\mathbf{Y}\mathbf{D}\mathbf{Y}^T \ .$$

$\mathbf{C}(0) \propto$ *covariance matrix* of the *resampled* data.

- kernel matrix $\mathbf{K} = \frac{1}{N}\mathbf{Y}^T\mathbf{Y}$

- Partition function

$$\begin{aligned} Z &= \int d^N\mathbf{x} \ \exp\left[-\frac{1}{2}\mathbf{x}^T\left(\mathbf{K}^{-1} + \mathbf{D}\right)\mathbf{x}\right] \\ &= |\mathbf{K}|^{\frac{1}{2}}\Gamma^{d/2}(2\pi)^{(N-d)/2}\int d^d\mathbf{z} \ \exp\left[-\frac{1}{2}\mathbf{z}^T\left(\mathbf{C}(\epsilon) + \Gamma\mathbf{I}\right)\mathbf{z}\right] \ . \end{aligned}$$

Z as generating function

$$
\begin{aligned}
-2\frac{\partial \ln Z}{\partial \epsilon}_{\epsilon=0} &= \frac{1}{\mu N}\sum_{j=1}^{N}\delta_{s_j,0}\,\mathrm{Tr}\,\mathbf{y}_j\mathbf{y}_j^T\mathbf{G}(\Gamma) \\
-2\frac{\partial \ln Z}{\partial \Gamma} &= \frac{d}{\Gamma}+\mathrm{Tr}\,\mathbf{G}(\Gamma)
\end{aligned}
$$

with

$$
\mathbf{G}(\Gamma) = (\mathbf{C}(0)+\Gamma\mathbf{I})^{-1} = \sum_k \frac{\mathbf{u}_k\mathbf{u}_k^T}{\lambda_k+\Gamma}
$$

Compare with (resample averaged) reconstruction error

$$
\mathcal{E}_r = \frac{1}{N_0}\mathbf{E}_\mathbf{D}\left[\sum_{\mathbf{y}_i\notin D;\lambda_l<\lambda}\mathrm{Tr}\left(\mathbf{y}_i\mathbf{y}_i^T\mathbf{u}_l\mathbf{u}_l^T\right)\right]
$$

Analytical Continuation

Reconstruction error

$$\mathcal{E}_r = \frac{1}{N_0} \mathbf{E}_\mathbf{D} \left[\sum_{\mathbf{y}_i \notin D; \lambda_l < \lambda} \mathrm{Tr} \left(\mathbf{y}_i \mathbf{y}_i^T \mathbf{u}_l \mathbf{u}_l^T \right) \right]$$

Use representation of the *Dirac* δ $\delta(x) = \lim_{\eta \to 0+} \Im \frac{1}{\pi(x - i\eta)}$ and get

$$\mathcal{E}_r = \mathcal{E}_r^0 + \int_{0+}^{\lambda} d\lambda' \, \varepsilon_r(\lambda')$$

where

$$\varepsilon_r(\lambda) = \frac{1}{\pi} \lim_{\eta \to 0+} \Im \frac{1}{N_0} \mathbf{E}_\mathbf{D} \left[\sum_j \delta_{s_j, 0} \, \mathrm{Tr} \left(\mathbf{y}_j \mathbf{y}_j^T \mathbf{G}(-\lambda - i\eta) \right) \right]$$

defines *error density* from all eigenvalues > 0 and and $\mathcal{E}_r^0$ is the contribution from eigenspace with $\lambda_k = 0$.

Replica Trick

Data averaged free energy

$$-\mathbf{E}_{\mathbf{D}}[\ln Z] = -\lim_{n \to 0} \frac{1}{n} \ln \mathbf{E}_{\mathbf{D}}[Z^n] \ ,$$

for integer n:

$$Z^{(n)} \doteq \mathbf{E}_{\mathbf{D}}[Z^n] = \int dx \ \psi_1(x) \ \psi_2(x)$$

where we set $x \doteq (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and

$$\psi_1(x) = \mathbf{E}_{\mathbf{D}} \left[\exp \left\{ -\frac{1}{2} \sum_{a=1}^{n} \mathbf{x}_a^T \mathbf{D} \mathbf{x}_a \right\} \right] \qquad \psi_2(x) = \exp \left[-\frac{1}{2} \sum_{a=1}^{n} \mathbf{x}_a^T \mathbf{K}^{-1} \mathbf{x}_a \right]$$

intractable!

Approximate Inference (EC: Opper & Winther)

$$p_1(x) = \frac{1}{Z_1}\psi_1(x)e^{-\Lambda_1 x^T x} \qquad p_0(x) = \frac{1}{Z_0}e^{-\frac{1}{2}\Lambda_0 x^T x}\ ,$$

with Λ_1 and Λ_0 "variational" parameters

$$\begin{aligned} Z^{(n)} &= Z_1 \int dx\, p_1(x)\, \psi_2(x)\, e^{\Lambda_1 x^T x} \\ &\approx Z_1 \int dx\, p_0(x)\, \psi_2(x)\, e^{\Lambda_1 x^T x} \equiv Z_{EC}^{(n)}(\Lambda_1, \Lambda_0) \end{aligned}$$

Match moments $\langle x^T x\rangle_1 = \langle x^T x\rangle_0$ & Stationarity w.r.t. Λ_1

<u>Final result</u>

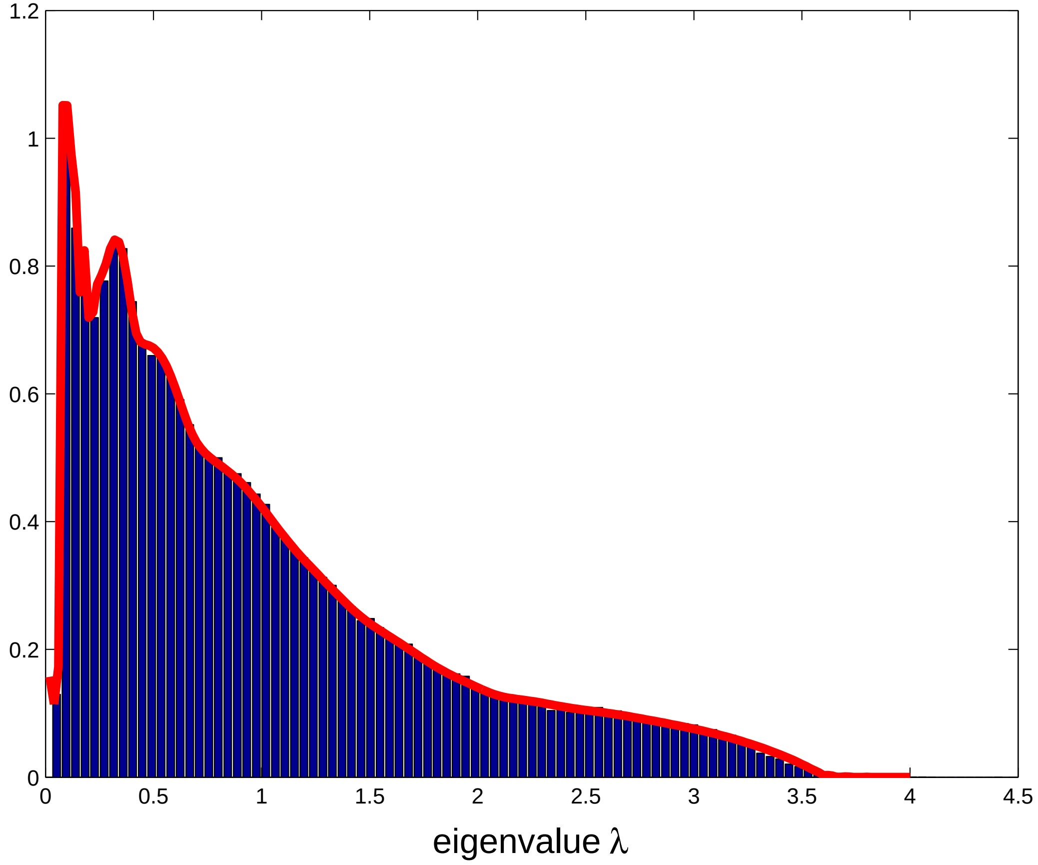
$$\begin{aligned} -\ln Z_{EC} &= -\mathbf{E}_{\mathbf{D}}\left[\ln \int d\mathbf{x}\, e^{-\frac{1}{2}\mathbf{x}^T(\mathbf{D}+(\Lambda_0-\Lambda)\mathbf{I})\mathbf{x}}\right] - \\ &\quad -\ln \int d\mathbf{x}\, e^{-\frac{1}{2}\mathbf{x}^T(\mathbf{K}^{-1}+\Lambda\mathbf{I})\mathbf{x}} + \ln \int d\mathbf{x}\, e^{-\frac{1}{2}\Lambda_0\mathbf{x}^T\mathbf{x}} \end{aligned}$$

where we have set $\Lambda = \Lambda_0 - \Lambda_1$. Tractable!

Result: Artificial Data

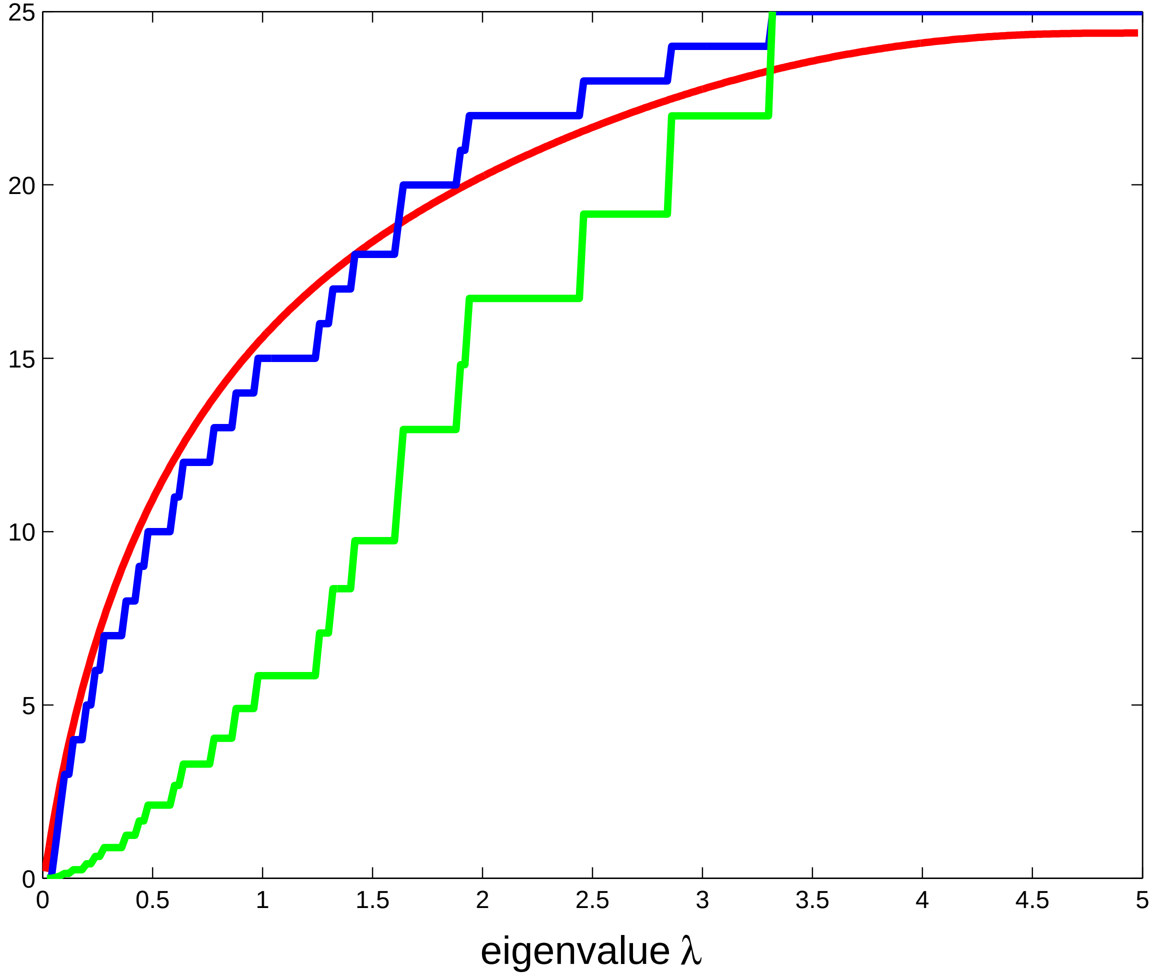
$N = 50$ data, Dim $= 25$, $3\times$ oversampled.

EC vs resampling



The PCA Reconstruction Error

(N = 32 artificial random data, Dim = 25) Approximate bootstrap 3$\times$ oversampled

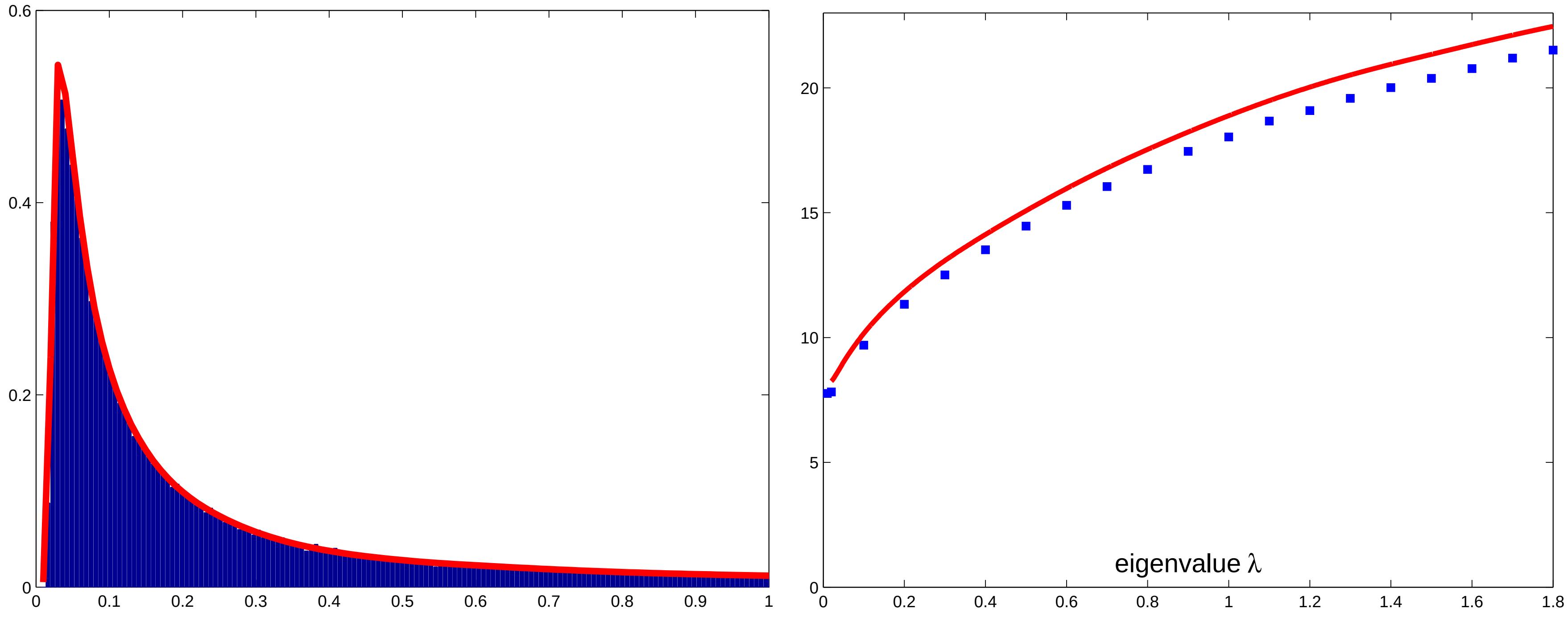


test error versus sum of eigenvalues (training error)

Approximate Bootstrap: handwritten Digits

($N = 100$ data, Dim $= 784$)

Density of eigenvalues and reconstruction error



The result without replicas

$$
\begin{aligned}
-\ln Z \;\; = \;\; & -\ln \int d\mathbf{x}\; e^{-\frac{1}{2}\mathbf{x}^T(\mathbf{D}+(\Lambda_0-\Lambda)\mathbf{I})\mathbf{x}} - \ln \int d\mathbf{x}\; e^{-\frac{1}{2}\mathbf{x}^T(\mathbf{K}^{-1}+\Lambda\mathbf{I})\mathbf{x}} + \\
& + \ln \int d\mathbf{x}\; e^{-\frac{1}{2}\Lambda_0\mathbf{x}^T\mathbf{x}} + \frac{1}{2}\ln\det(\mathbf{I}+\mathbf{r})
\end{aligned}
$$

with

$$
\mathbf{r}_{ij} = \left(1 - \frac{\Lambda_0}{\Lambda_0 - \Lambda + D_i}\right)\left(\Lambda_0\left(\mathbf{K}^{-1} + \Lambda\mathbf{I}\right)^{-1} - \mathbf{I}\right)_{ij} \quad .
$$

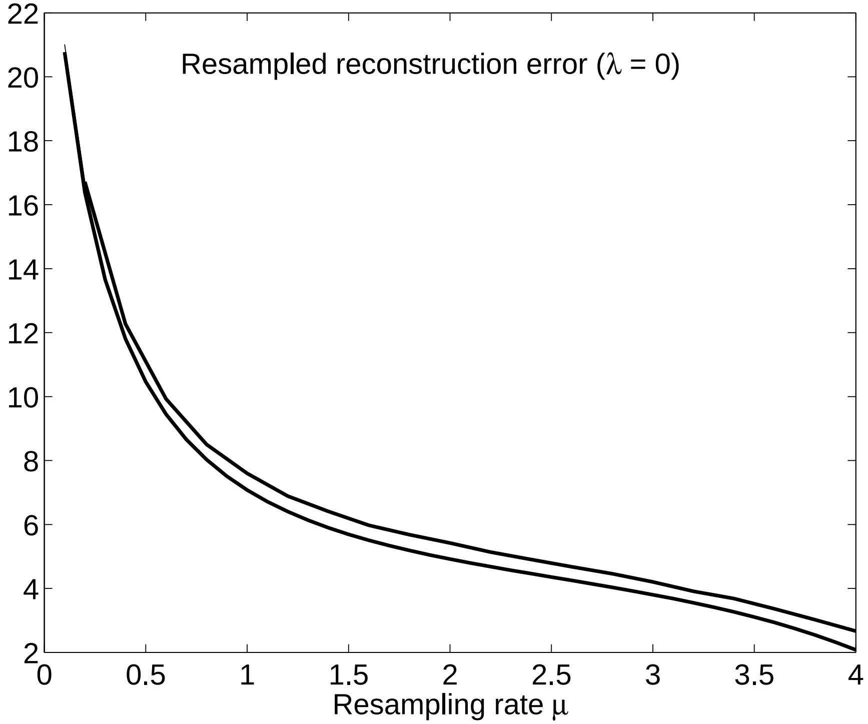
Expand

$$
\ln\det\left(\mathbf{I}+\mathbf{r}\right) = \mathrm{Tr}\ln\left(\mathbf{I}+\mathbf{r}\right) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}\,\mathrm{Tr}\left(\mathbf{r}^k\right)
$$

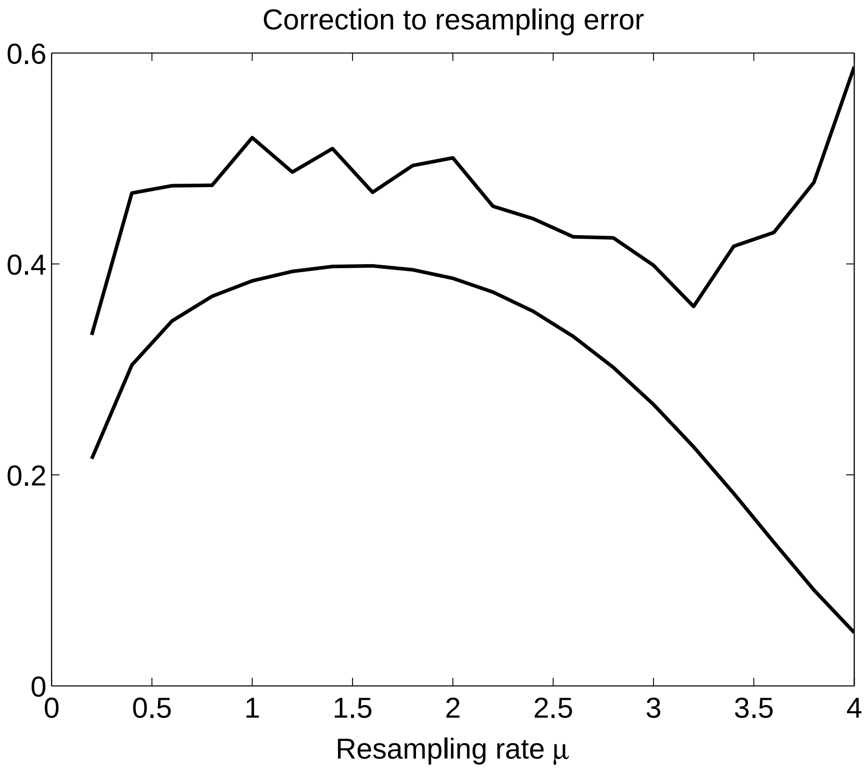
We have $\mathbf{E}_{\mathbf{D}}[\mathbf{r}_{ij}] = 0 \rightarrow$ 1.order term vanishes after average, 2.order yields on average

$$
\Delta F = -\frac{1}{4}\sum_i \left(\Lambda_0\left(\mathbf{K}^{-1} + \Lambda\mathbf{I}\right)^{-1}_{ii} - 1\right)^2 \times \sum_i \mathbf{E}_{\mathbf{D}}\left(\frac{\Lambda_0}{\Lambda_0 - \Lambda + D_i} - 1\right)^2
$$

Correction



Resampled reconstruction error (λ = 0)
Resampling rate μ
22
20
18
16
14
12
10
8
6
4
2
0
0.5
1
1.5
2
2.5
3
3.5
4



Correction to resampling error
Resampling rate μ
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
3.5
4

Correction to EC

$$\frac{Z^{(n)}}{Z_1} = \int dx p_1(x)\, \psi_2(x)\, e^{\Lambda_1 x^T x} = \int dx \psi_2(x)\, e^{\frac{1}{2}\Lambda x^T x} \left\{ \int \frac{dk}{(2\pi)^{Nn}} e^{-ik^T x} \chi(k) \right\}$$

where $\chi(k) \doteq \int dx\; p_1(x)\; e^{-ik^T x}$ is the *characteristic function* of the density p_1.

Cumulant expansion starts with a quadratic term (EC)

$$\ln \chi(k) = -\frac{M_2}{2} k^T k + R(k)\ , \qquad (1)$$

where $M_2 = \langle \mathbf{x}_a^T \mathbf{x}_a \rangle_1$.

Expand 4-th order term in $R(k)$ as $e^{R(k)} = 1 + R(k) + \ldots$ leads to ΔF.

Possibility of perturbative improvement?

Conclusion

- Non–Bayesian inference problems can be related to “hidden” probabilistic models via analytic continuation.

- EC approximate inference appears to be robust and survives analytic continuation and limits.