Journal of Machine Learning Research 9 (2008) 2035-2078

Approximations for Binary Gaussian Process Classification

Hannes Nickisch and Carl Edward Rasmussen

Presented by Shaobo Han, Duke University
April 19, 2013

Outline

1. Binary Gaussian Process Classification
2. Gaussian Approximation Methods
   - Laplace Approximation (LA)
   - Expectation Propagation (EP)
   - KL-Divergence Minimization (KL)
   - Variational Bound (VB)
3. Experimental Results
   - Highly close-to-Gaussian Posterior
   - Highly non-Gaussian Posterior

Gaussian Process for Binary Classification

Factorial Likelihood:

$$\mathbb{P}(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} \mathbb{P}(y_i|f_i) = \prod_{i=1}^{n} \mathrm{sig}(y_i f_i), \quad y_i \in \{-1,+1\}, \quad \mathrm{sig}: \mathbb{R} \to [0,1] \quad (1)$$

where $\mathrm{sig}_{\mathrm{logit}}(t) := 1/(1+e^{-t})$, $\mathrm{sig}_{\mathrm{probit}}(t) := \int_{-\infty}^{t} \mathcal{N}(\tau|0,1)d\tau$

Non-Gaussian Posterior:

$$\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}) = \frac{\mathbb{P}(\mathbf{y}|\mathbf{f})\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})}{\mathbb{P}(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})} = \frac{\prod_{i=1}^{n} \mathrm{sig}(y_i f_i)\mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})}{\int \mathbb{P}(\mathbf{y}|\mathbf{f})\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})d\mathbf{f}} \quad (2)$$

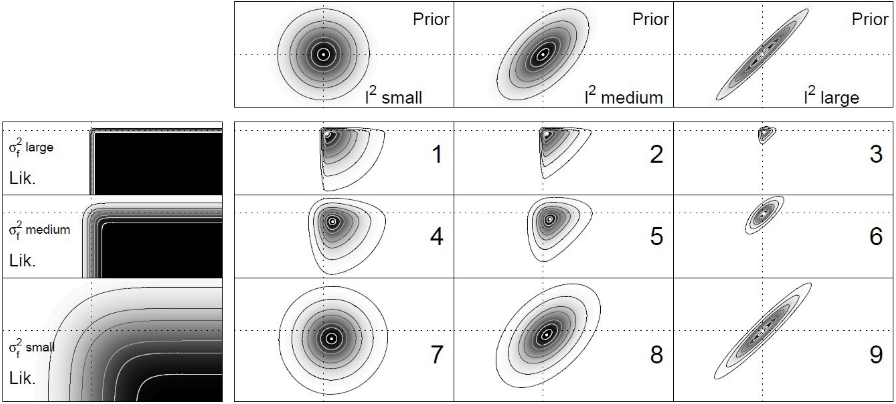
Prediction:

$$\begin{aligned} \mathbb{P}(f_*|\mathbf{x}_*,\mathbf{y},\mathbf{X},\boldsymbol{\theta}) &= \int \mathbb{P}(f_*|\mathbf{f},\mathbf{x}_*,\mathbf{X},\boldsymbol{\theta})\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})d\mathbf{f} \\ \mathbb{P}(y_*|\mathbf{x}_*,\mathbf{y},\mathbf{X},\boldsymbol{\theta}) &= \int \mathrm{sig}(y_* f_*)\mathbb{P}(f_*|\mathbf{x}_*,\mathbf{y},\mathbf{X},\boldsymbol{\theta})df_* \quad (3) \end{aligned}$$

Stationary Covariance Functions:

$$\mathbf{K} = \mathbf{k}(\mathbf{x},\mathbf{x}',\boldsymbol{\theta}) = \sigma_f^2 g(|\mathbf{x}-\mathbf{x}'|/l), \quad g: \mathbb{R} \to \mathbb{R}, \quad \boldsymbol{\theta} = \{\sigma_f, l\} \quad (4)$$

Prior, Likelihood, and Exact Posterior



$$\lim_{l \to 0} \mathbf{K} = \sigma_f^2 \mathbf{I}, \quad \lim_{l \to \infty} \mathbf{K} = \sigma_f^2 \mathbf{1}\mathbf{1}^T, \quad \lim_{\sigma_f \to 0} \mathrm{sig}(t) = 0.5 \tag{5}$$

$$\lim_{\sigma_f \to \infty} \mathrm{sig}(t) = \mathrm{step}(t) := \{0, t < 0; 0.5, t = 0; 1, 0 < t\} \tag{6}$$

Gaussian Approximation & Effective Likelihood

Approximate Gaussian Posteriors (log-concave → unimodality):

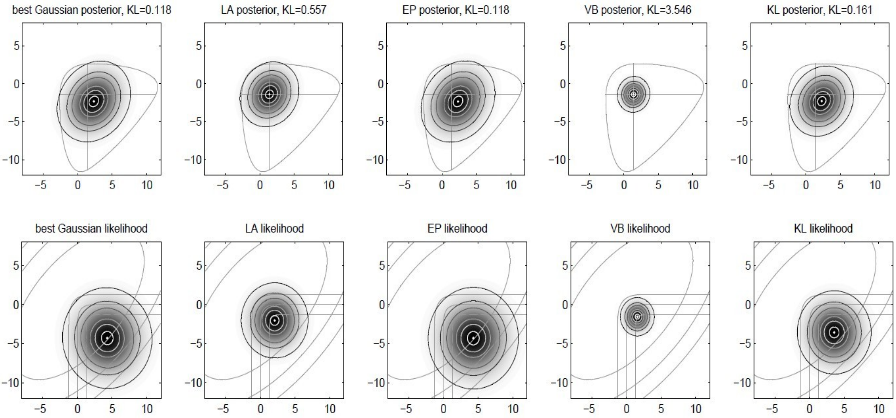
$$
\begin{aligned}
\ln \mathbb{P}(\mathbf{f}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) &= -\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f} + \sum_{i=1}^{n} \ln \mathbb{P}(y_i|f_i) + \mathrm{const}_{\mathbf{f}} \\
&\approx -\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}\mathbf{f}^T\mathbf{W}\mathbf{f} + \mathbf{b}^T\mathbf{f} + \mathrm{const}_{\mathbf{f}} \\
&= -\frac{1}{2}(\mathbf{f}-\mathbf{m})^T(\mathbf{K}^{-1}+\mathbf{W})(\mathbf{f}-\mathbf{m}) + \mathrm{const}_{\mathbf{f}} \\
&= \ln \mathcal{N}(\mathbf{f}|\mathbf{m}, \mathbf{V}) := \ln \mathbb{Q}(\mathbf{f}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta})
\end{aligned} \tag{7}
$$

where $\mathbf{m} = (\mathbf{K}^{-1}+\mathbf{W})^{-1}\mathbf{b}$, $\mathbf{V}^{-1} = \mathbf{K}^{-1}+\mathbf{W}$.
Effective Likelihood (Gaussian factor):

$$
\mathbb{Q}(\mathbf{y}|\mathbf{f}) \propto \frac{\mathcal{N}(\mathbf{f}|\mathbf{m}, \mathbf{V})}{\mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})} \propto \mathcal{N}\left(\mathbf{f}|(\mathbf{K}\mathbf{W})^{-1}\mathbf{m}+\mathbf{m}, \mathbf{W}^{-1}\right) \tag{8}
$$

such that $\mathbb{Q}(\mathbf{f}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) \propto \mathbb{Q}(\mathbf{y}|\mathbf{f})\mathbb{P}(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})$

Gaussian Approximation Methods



Laplace Approximation (LA) (Williams and Barber, 1998)
Expectation Propagation (EP) (Minka, 2001a)
Kullback-Leibler divergence (KL) minimization (Opper and Archambeau, 2008) comprising Variational Bounding (VB) (Gibbs and Mackay, 2000)

Log Marginal Likelihood

Agreement of model and observed data is typically measured by the marginal likelihood Z

$$
\begin{aligned}
\ln Z &= \ln \mathbb{P}(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = \ln \int \mathbb{P}(\mathbf{y}|\mathbf{f})\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})d\mathbf{f} \\
&= \ln \int \mathbb{Q}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})\frac{\mathbb{P}(\mathbf{y}|\mathbf{f})\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})}{\mathbb{Q}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})}d\mathbf{f} \\
&\geq \int \mathbb{Q}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}) \ln \frac{\mathbb{P}(\mathbf{y}|\mathbf{f})\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})}{\mathbb{Q}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})}d\mathbf{f} \\
&=: \ln Z_B = \ln Z - \mathrm{KL}(\mathbb{Q}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})||\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})) \qquad (9)
\end{aligned}
$$

Accurate marginal likelihood estimates Z are a key to hyperparameter learning. For example, model selection by type II maximum likelihood also known as the evidence framework (MacKay, 1992)

Marginal likelihood (evidence) $\mathbb{P}(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})$ is approximated in Laplace Approximation (LA) and Expectation Propagation (EP)

Laplace Approximation (LA)

Posterior:

$$
\begin{aligned}
\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}) &\approx \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V}) = \mathcal{N}\left(\mathbf{f}|\mathbf{m},(\mathbf{K}^{-1}+\mathbf{W})^{-1}\right) \\
\mathbf{m} &= \arg\max_{\mathbf{f}\in\mathbb{R}^n} \mathbb{P}(\mathbf{y}|\mathbf{f})\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta}) \\
\mathbf{W} &= -\left.\frac{\partial^2 \ln \mathbb{P}(\mathbf{y}|\mathbf{f})}{\partial\mathbf{f}\partial\mathbf{f}^T}\right|_{\mathbf{f}=\mathbf{m}} = -\left[\left.\frac{\partial \ln \mathbb{P}(y_i|f_i)}{\partial f_i^2}\right|_{f_i=m_i}\right]_{ii} \qquad (10)
\end{aligned}
$$

Log Marginal Likelihood:

Define $\boldsymbol{\Psi}(\mathbf{f}) := \ln \mathbb{P}(\mathbf{y}|\mathbf{f}) + \ln \mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})$, a Taylor expansion of $\boldsymbol{\Psi}$ is then given by $\boldsymbol{\Psi}(\mathbf{f}) \approx \boldsymbol{\Psi}(\mathbf{m}) - \frac{1}{2}(\mathbf{f}-\mathbf{m})^T(\mathbf{K}^{-1}+\mathbf{W})(\mathbf{f}-\mathbf{m})$

$$
\begin{aligned}
\ln Z &= \ln \mathbb{P}(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = \ln \int \mathbb{P}(\mathbf{y}|\mathbf{f})\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})d\mathbf{f} = \ln \int \exp\left(\boldsymbol{\Psi}(\mathbf{f})\right)d\mathbf{f} \\
&\approx \boldsymbol{\Psi}(\mathbf{m}) + \ln \int \exp\left(-\frac{1}{2}(\mathbf{f}-\mathbf{m})^T(\mathbf{K}^{-1}+\mathbf{W})(\mathbf{f}-\mathbf{m})\right)d\mathbf{f} \\
&= \ln \mathbb{P}(\mathbf{y}|\mathbf{m}) - \frac{1}{2}\mathbf{m}^T\mathbf{K}^{-1}\mathbf{m} + \frac{1}{2}\ln|\mathbf{I}+\mathbf{K}\mathbf{W}| \qquad (11)
\end{aligned}
$$

Key Idea: Quadratic expansion around the mode

Expectation Propagation (EP)

Posterior:

$$\begin{aligned}\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}) &\approx \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V}) = \mathcal{N}\left(\mathbf{f}|\mathbf{m},(\mathbf{K}^{-1}+\mathbf{W})^{-1}\right) \\ \mathbf{W} &= [\sigma_i^{-2}]_{ii}, \quad \boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)^T \\ \mathbf{m} &= \mathbf{V}\mathbf{W}\boldsymbol{\mu} = [\mathbf{I}-\mathbf{K}(\mathbf{K}+\mathbf{W}^{-1})^{-1}]\mathbf{K}\mathbf{W}\boldsymbol{\mu} \end{aligned} \tag{12}$$

Log Marginal Likelihood:

$$\begin{aligned}\ln Z &= \ln \mathbb{P}(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = \ln \int \prod_{i=1}^{n} \mathbb{P}(y_i|f_i)\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})d\mathbf{f} \\ &\approx \ln \int \prod_{i=1}^{n} t_i(f_i,\mu_i,\sigma_i^2,Z_i)\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})d\mathbf{f} = \ln Z_{EP}\end{aligned} \tag{13}$$

Key Idea: Iteratively matching marginal moments (μ_i, σ_i^2, Z_i) between $\mathbb{Q}(f_i) := \int \mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K}) \prod_{j=1}^{n} Z_j \mathcal{N}(f_j|\mu_j,\sigma_j^2)df_{\neg i}$ (approximate marginal posteriors) and $\int \mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})\mathbb{P}(y_i|f_i) \prod_{j\neq i}^{n} Z_j \mathcal{N}(f_j|\mu_j,\sigma_j^2)df_{\neg i}$ based on the exact likelihood term $\mathbb{P}(y_i|f_i)$

KL-Divergence Minimization (KL)

Posterior:

$$\begin{aligned}\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}) &\approx \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V}) = \mathcal{N}\left(\mathbf{f}|\mathbf{m},(\mathbf{K}^{-1}+\mathbf{W})^{-1}\right)\\ \mathbf{W} &= -2\boldsymbol{\Lambda}, \quad \mathbf{m} = \mathbf{K}\boldsymbol{\alpha}\end{aligned} \tag{14}$$

Log Marginal Likelihood: $\ln Z_B = \ln Z - \mathrm{KL}(\mathbb{Q}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})||\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}))$

Key Idea:

$$\begin{aligned}\mathrm{KL}(\mathbb{Q}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})||\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})) &= \int \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V}) \ln \frac{\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})}{\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})} d\mathbf{f}\\ &= a(\mathbf{m},\mathbf{V}) - \frac{1}{2}\ln|\mathbf{V}| + \frac{1}{2}\mathbf{m}^T\mathbf{K}^{-1}\mathbf{m} + \frac{1}{2}\mathrm{tr}(\mathbf{K}^{-1}\mathbf{V})\end{aligned} \tag{15}$$

where $a(\mathbf{m},\mathbf{V}) = -\int \mathcal{N}(\mathbf{f}) \left[\sum_{i=1}^n \ln \mathrm{sig}(\sqrt{v_{ii}}y_i f_i + m_i y_i)\right] d\mathbf{f}$

$$\frac{\partial \mathrm{KL}}{\partial \mathbf{m}} = \frac{\partial a}{\partial \mathbf{m}} - \mathbf{K}^{-1}\mathbf{m} = \mathbf{0}, \quad \frac{\partial \mathrm{KL}}{\partial \mathbf{V}} = \frac{\partial a}{\partial \mathbf{V}} + \frac{1}{2}\mathbf{V}^{-1} - \frac{1}{2}\mathbf{K}^{-1} = \mathbf{0} \tag{16}$$

$$(\mathbf{m},\mathbf{V}) \mapsto [\boldsymbol{\alpha},\boldsymbol{\Lambda}_{ii}], \quad \mathcal{O}(n^2) \mapsto \mathcal{O}(n), \quad \boldsymbol{\alpha} = \frac{\partial a}{\partial \mathbf{m}} = \mathbf{K}^{-1}\mathbf{m}, \quad \boldsymbol{\Lambda} = \frac{\partial a}{\partial \mathbf{V}} \tag{17}$$

Variational Bound (VB)

Posterior:

$$\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\theta) \approx \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V}) = \mathcal{N}\left(\mathbf{f}|\mathbf{m},(\mathbf{K}^{-1}+\mathbf{W})^{-1}\right)$$
$$\mathbf{W} = -2\mathbf{A}_\varsigma, \quad \mathbf{m} = \mathbf{V}(\mathbf{y}\odot\mathbf{b}_\varsigma) = (\mathbf{K}^{-1}-2\mathbf{A}_\varsigma)^{-1}(\mathbf{y}\odot\mathbf{b}_\varsigma) \tag{18}$$

Log Marginal Likelihood:

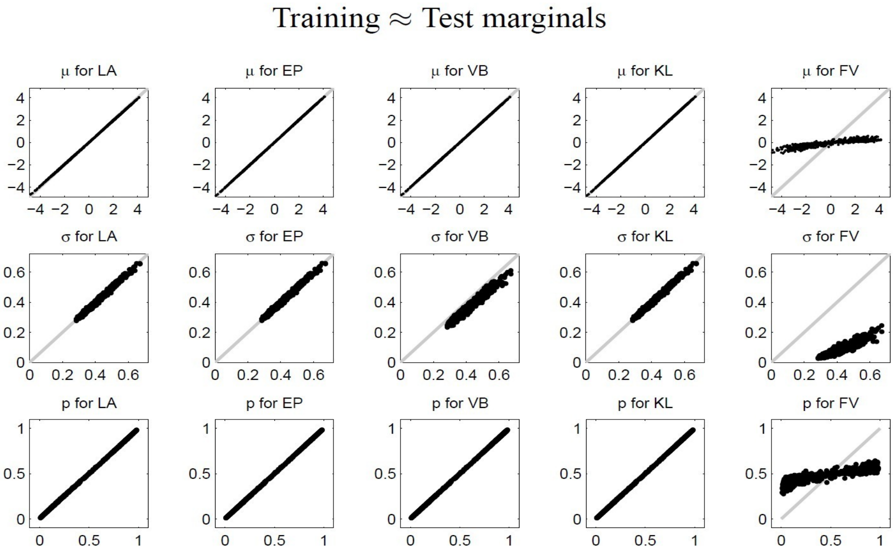
$$\ln Z_{VB} = \mathbf{c}^T\mathbf{1} + \frac{1}{2}(\mathbf{b}\odot\mathbf{y})^T(\mathbf{K}^{-1}-2\mathbf{A})^{-1}(\mathbf{b}\odot\mathbf{y}) - \frac{1}{2}\ln|\mathbf{I}-2\mathbf{A}\mathbf{K}| \tag{19}$$

Key Idea (Individual likelihood bounds):

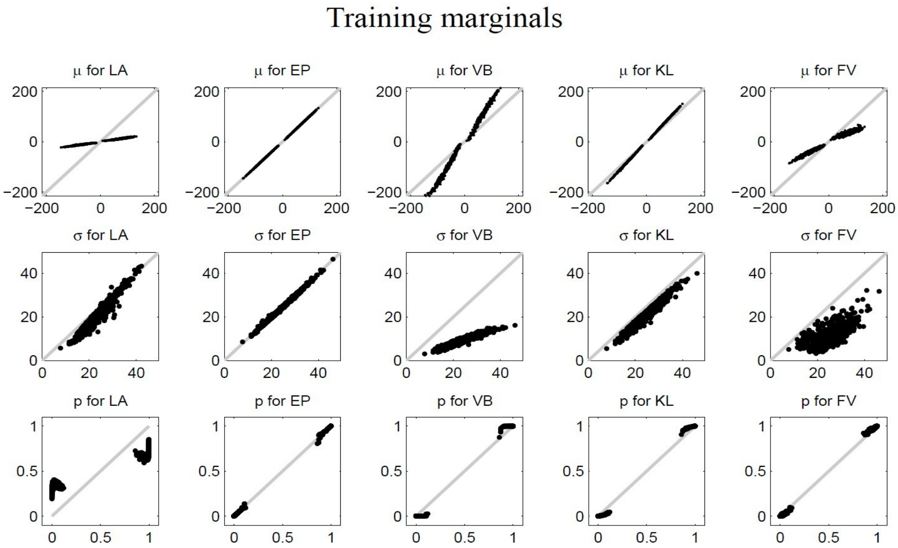
$$\begin{aligned}
\mathbb{P}(y_i|f_i) &\geq \exp\left(a_i f_i^2 + b_i y_i f_i + c_i\right), \quad \forall\ f_i \in \mathbb{R}\ \forall\ i \\
\mathbb{P}(\mathbf{y}|\mathbf{f}) &\geq \exp\left(\mathbf{f}^T\mathbf{A}\mathbf{f} + (\mathbf{b}\odot\mathbf{y})^T\mathbf{f} + \mathbf{c}^T\mathbf{1}\right) =: \mathbb{Q}(y|\mathbf{f},\mathbf{A},\mathbf{b},\mathbf{c}) \ \ \forall \mathbf{f}\in\mathbb{R}^n \\
Z &= \int \mathbb{P}(\mathbf{f}|\mathbf{X})\mathbb{P}(\mathbf{y}|\mathbf{f})d\mathbf{f} \geq \int \mathbb{P}(\mathbf{f}|\mathbf{X})\mathbb{Q}(\mathbf{y}|\mathbf{f},\mathbf{A},\mathbf{b},\mathbf{c})d\mathbf{f} = Z_{VB}
\end{aligned} \tag{20}$$

$$(\mathbf{A},\mathbf{b},\mathbf{c}) \mapsto \varsigma \mapsto (\mathbf{m}_\varsigma,\mathbf{V}_\varsigma), \quad \mathbf{Z} \geq \mathbf{Z}_B \geq \mathbf{Z}_{VB}, \quad \mathbf{Z}_{EP} \geq \mathbf{Z}_B \tag{21}$$

USPS 3 vs.5: Highly close-to-Gaussian Posterior



USPS 3 vs.5: Highly non-Gaussian Posterior



USPS 3 vs.5: Highly non-Gaussian Posterior