Hierarchic Bayesian Inference of Mixed Modality Brain Imaging for Clinical Diagnostics

Mark Girolami

Department of Statistics
University of Warwick

CRiSM Workshop
Statistical Challenges in Neuroscience

September 3 2014

Scope of Work and Collaboration

- Ongoing collaborative study of neuroimaging as clinical diagnostic aid

Scope of Work and Collaboration

- Ongoing collaborative study of neuroimaging as clinical diagnostic aid
- Institute of Psychiatry, Kings College London

Scope of Work and Collaboration

- Ongoing collaborative study of neuroimaging as clinical diagnostic aid
- Institute of Psychiatry, Kings College London
- Brighton and Sussex Medical School

Scope of Work and Collaboration

- Ongoing collaborative study of neuroimaging as clinical diagnostic aid
- Institute of Psychiatry, Kings College London
- Brighton and Sussex Medical School
- Welcome Trust Centre for Neuroimaging, University College London

Scope of Work and Collaboration

- Ongoing collaborative study of neuroimaging as clinical diagnostic aid
- Institute of Psychiatry, Kings College London
- Brighton and Sussex Medical School
- Welcome Trust Centre for Neuroimaging, University College London
- Early stages of neuro-degenerative diseases clinically indistinguishable

Scope of Work and Collaboration

- Ongoing collaborative study of neuroimaging as clinical diagnostic aid
- Institute of Psychiatry, Kings College London
- Brighton and Sussex Medical School
- Welcome Trust Centre for Neuroimaging, University College London
- Early stages of neuro-degenerative diseases clinically indistinguishable
- Can structural neuroimaging assist in distinguishing early stage disease?

Parkinsonian Type Disorders

- Progressive Supranuclear Palsy (PSP)

Parkinsonian Type Disorders

- Progressive Supranuclear Palsy (PSP)
- Multiple System Atrophy (MSA)

Parkinsonian Type Disorders

- Progressive Supranuclear Palsy (PSP)
- Multiple System Atrophy (MSA)
- Idiopathic Parkinson's Disease (IPD)

Parkinsonian Type Disorders

- Progressive Supranuclear Palsy (PSP)
- Multiple System Atrophy (MSA)
- Idiopathic Parkinson's Disease (IPD)
- Clinically indistinguishable in early stages

Parkinsonian Type Disorders

- Progressive Supranuclear Palsy (PSP)
- Multiple System Atrophy (MSA)
- Idiopathic Parkinson's Disease (IPD)
- Clinically indistinguishable in early stages
- Different prognoses - PSP & MSA relentless progression, IPD no substantial reduction in life expectancy

Parkinsonian Type Disorders

- Progressive Supranuclear Palsy (PSP)
- Multiple System Atrophy (MSA)
- Idiopathic Parkinson's Disease (IPD)
- Clinically indistinguishable in early stages
- Different prognoses - PSP & MSA relentless progression, IPD no substantial reduction in life expectancy
- Different responses to treatment - IPD good response to dopamine therapy, PSP & MSA poor response

Parkinsonian Type Disorders

- Progressive Supranuclear Palsy (PSP)
- Multiple System Atrophy (MSA)
- Idiopathic Parkinson's Disease (IPD)
- Clinically indistinguishable in early stages
- Different prognoses - PSP & MSA relentless progression, IPD no substantial reduction in life expectancy
- Different responses to treatment - IPD good response to dopamine therapy, PSP & MSA poor response
- Objective biomarkers predictive of early disease state useful in reducing clinical trial misdiagnosis

Parkinsonian Type Disorders

- Progressive Supranuclear Palsy (PSP)
- Multiple System Atrophy (MSA)
- Idiopathic Parkinson's Disease (IPD)
- Clinically indistinguishable in early stages
- Different prognoses - PSP & MSA relentless progression, IPD no substantial reduction in life expectancy
- Different responses to treatment - IPD good response to dopamine therapy, PSP & MSA poor response
- Objective biomarkers predictive of early disease state useful in reducing clinical trial misdiagnosis

MRI Objective Diagnostic Markers

- No published studies demonstrate automated approach of individual diagnostics clinically useful

MRI Objective Diagnostic Markers

- No published studies demonstrate automated approach of individual diagnostics clinically useful..... until now

MRI Objective Diagnostic Markers

- No published studies demonstrate automated approach of individual diagnostics clinically useful..... until now..... but cautious progress

MRI Objective Diagnostic Markers

- No published studies demonstrate automated approach of individual diagnostics clinically useful..... until now..... but cautious progress
- Existing studies employ manual measurements (radiological and voxel-based morphometry) from MRI

MRI Objective Diagnostic Markers

- No published studies demonstrate automated approach of individual diagnostics clinically useful..... until now..... but cautious progress
- Existing studies employ manual measurements (radiological and voxel-based morphometry) from MRI
- Radiological MRI operator-dependent & time consuming, insufficient specificity for MSA and PSP

MRI Objective Diagnostic Markers

- No published studies demonstrate automated approach of individual diagnostics clinically useful..... until now..... but cautious progress
- Existing studies employ manual measurements (radiological and voxel-based morphometry) from MRI
- Radiological MRI operator-dependent & time consuming, insufficient specificity for MSA and PSP
- Voxel morphometry limited ability to predict disease state at individual level

MRI Objective Diagnostic Markers

- No published studies demonstrate automated approach of individual diagnostics clinically useful..... until now..... but cautious progress
- Existing studies employ manual measurements (radiological and voxel-based morphometry) from MRI
- Radiological MRI operator-dependent & time consuming, insufficient specificity for MSA and PSP
- Voxel morphometry limited ability to predict disease state at individual level
- Previous studies at single subject level using statistical discriminant analysis unable to accurately discriminate all diagnostic groups for Parkinsonian disorders

Discrimination via Anatomical Network Patterns of Brain Regions

- Preliminary assessment of network patterns of brain regions for discrimination of Parkinsonian disorders

Discrimination via Anatomical Network Patterns of Brain Regions

- Preliminary assessment of network patterns of brain regions for discrimination of Parkinsonian disorders
- Networks of subcortical regions defined based on known distribution of PSP or MSA/IPD pathology used to test working hypothesis

Discrimination via Anatomical Network Patterns of Brain Regions

- Preliminary assessment of network patterns of brain regions for discrimination of Parkinsonian disorders
- Networks of subcortical regions defined based on known distribution of PSP or MSA/IPD pathology used to test working hypothesis
- Employ discriminant analysis (GP) to assess diagnostic capability of full network

Discrimination via Anatomical Network Patterns of Brain Regions

- Preliminary assessment of network patterns of brain regions for discrimination of Parkinsonian disorders
- Networks of subcortical regions defined based on known distribution of PSP or MSA/IPD pathology used to test working hypothesis
- Employ discriminant analysis (GP) to assess diagnostic capability of full network
- Can MSA subtypes (P & C) also be discriminated given different burdens on brainstem and ganglia pathology

Discrimination via Anatomical Network Patterns of Brain Regions

- Preliminary assessment of network patterns of brain regions for discrimination of Parkinsonian disorders
- Networks of subcortical regions defined based on known distribution of PSP or MSA/IPD pathology used to test working hypothesis
- Employ discriminant analysis (GP) to assess diagnostic capability of full network
- Can MSA subtypes (P & C) also be discriminated given different burdens on brainstem and ganglia pathology
- Can combination of network components (e.g. midbrain, brainstem, cerebellar peduncle, etc) outperform whole-brain approach

Discrimination via Anatomical Network Patterns of Brain Regions

- Preliminary assessment of network patterns of brain regions for discrimination of Parkinsonian disorders
- Networks of subcortical regions defined based on known distribution of PSP or MSA/IPD pathology used to test working hypothesis
- Employ discriminant analysis (GP) to assess diagnostic capability of full network
- Can MSA subtypes (P & C) also be discriminated given different burdens on brainstem and ganglia pathology
- Can combination of network components (e.g. midbrain, brainstem, cerebellar peduncle, etc) outperform whole-brain approach important as cortical atrophy reported in all disorders

Case Selection

- PSP - 17 cases, MSA - 19 cases, IPD - 14 cases.

Case Selection

- PSP - 17 cases, MSA - 19 cases, IPD - 14 cases. All diagnosed with established criteria

Case Selection

- PSP - 17 cases, MSA - 19 cases, IPD - 14 cases. All diagnosed with established criteria though limited confirmation pathology - 8 patients

Case Selection

- PSP - 17 cases, MSA - 19 cases, IPD - 14 cases. All diagnosed with established criteria though limited confirmation pathology - 8 patients
- Small cohort study

Case Selection

- PSP - 17 cases, MSA - 19 cases, IPD - 14 cases. All diagnosed with established criteria though limited confirmation pathology - 8 patients
- Small cohort study first demonstration of feasibility at individual level of discrimination of disorders

Case Selection

- PSP - 17 cases, MSA - 19 cases, IPD - 14 cases. All diagnosed with established criteria though limited confirmation pathology - 8 patients
- Small cohort study first demonstration of feasibility at individual level of discrimination of disorders
- Caution required interpreting reported results and extrapolating

Case Selection

- PSP - 17 cases, MSA - 19 cases, IPD - 14 cases. All diagnosed with established criteria though limited confirmation pathology - 8 patients
- Small cohort study first demonstration of feasibility at individual level of discrimination of disorders
- Caution required interpreting reported results and extrapolating
- Interpretation of combination of network components in discrimination consistent with known pathology

Case Selection

- PSP - 17 cases, MSA - 19 cases, IPD - 14 cases. All diagnosed with established criteria though limited confirmation pathology - 8 patients
- Small cohort study first demonstration of feasibility at individual level of discrimination of disorders
- Caution required interpreting reported results and extrapolating
- Interpretation of combination of network components in discrimination consistent with known pathology very important in discriminating MSA subtypes - combination of cerebellum, brainstem and putamen.

Case Selection

- PSP - 17 cases, MSA - 19 cases, IPD - 14 cases. All diagnosed with established criteria though limited confirmation pathology - 8 patients
- Small cohort study first demonstration of feasibility at individual level of discrimination of disorders
- Caution required interpreting reported results and extrapolating
- Interpretation of combination of network components in discrimination consistent with known pathology very important in discriminating MSA subtypes - combination of cerebellum, brainstem and putamen.
- Encouraging start and ongoing clinical studies in progress

Bayesian Hierarchic Model

- Choice of nonparametric Bayesian model for posterior predictive label value

Bayesian Hierarchic Model

- Choice of nonparametric Bayesian model for posterior predictive label value
- Due to small cohort - no Big Data here in such medical studies

Bayesian Hierarchic Model

- Choice of nonparametric Bayesian model for posterior predictive label value
- Due to small cohort - no Big Data here in such medical studies
- Gaussian Process functional prior - well studied in ML and Comp Stats literature

Bayesian Hierarchic Model

- Choice of nonparametric Bayesian model for posterior predictive label value
- Due to small cohort - no Big Data here in such medical studies
- Gaussian Process functional prior - well studied in ML and Comp Stats literature
- Challenge to perform Bayesian marginalisation 'exactly'

Bayesian Hierarchic Model

- Choice of nonparametric Bayesian model for posterior predictive label value
- Due to small cohort - no Big Data here in such medical studies
- Gaussian Process functional prior - well studied in ML and Comp Stats literature
- Challenge to perform Bayesian marginalisation 'exactly'
- Poor mixing of variables in top level of the hierarchy well known issue

Bayesian Hierarchic Model

- Choice of nonparametric Bayesian model for posterior predictive label value
- Due to small cohort - no Big Data here in such medical studies
- Gaussian Process functional prior - well studied in ML and Comp Stats literature
- Challenge to perform Bayesian marginalisation 'exactly'
- Poor mixing of variables in top level of the hierarchy well known issue
- Consider novel and general solution exploiting approximation schemes

Bayesian Hierarchic GP Model

- Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n input vectors described by d covariates and associated with observed univariate responses $\mathbf{y} = \{y_1, \ldots, y_n\}$ with $y_i \in \{-1, +1\}$.

Bayesian Hierarchic GP Model

- Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n input vectors described by d covariates and associated with observed univariate responses $\mathbf{y} = \{y_1, \ldots, y_n\}$ with $y_i \in \{-1, +1\}$.
- Let $\mathbf{f} = \{f_1, \ldots, f_n\}$ be a set of latent variables. Assume that the class labels have a Bernoulli distribution with success probability given by a transformation of the latent variables:

$$p(y_i|f_i) = \Phi(y_i f_i). \tag{1}$$

Bayesian Hierarchic GP Model

- Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n input vectors described by d covariates and associated with observed univariate responses $\mathbf{y} = \{y_1, \ldots, y_n\}$ with $y_i \in \{-1, +1\}$.
- Let $\mathbf{f} = \{f_1, \ldots, f_n\}$ be a set of latent variables. Assume that the class labels have a Bernoulli distribution with success probability given by a transformation of the latent variables:

$$p(y_i|f_i) = \Phi(y_i f_i). \tag{1}$$

Here Φ denotes the cumulative function of the Gaussian density; based on this modeling assumption, the likelihood function is:

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i). \tag{2}$$

Bayesian Hierarchic GP Model

- Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n input vectors described by d covariates and associated with observed univariate responses $\mathbf{y} = \{y_1, \ldots, y_n\}$ with $y_i \in \{-1, +1\}$.
- Let $\mathbf{f} = \{f_1, \ldots, f_n\}$ be a set of latent variables. Assume that the class labels have a Bernoulli distribution with success probability given by a transformation of the latent variables:

$$p(y_i|f_i) = \Phi(y_i f_i). \tag{1}$$

Here Φ denotes the cumulative function of the Gaussian density; based on this modeling assumption, the likelihood function is:

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i). \tag{2}$$

The latent variables $\mathbf{f}$ are given a zero mean GP prior with covariance K:

$$\mathbf{f} \sim \mathcal{N}(\mathbf{f}|\mathbf{0}, K). \tag{3}$$

Bayesian Hierarchic GP Model

- Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n input vectors described by d covariates and associated with observed univariate responses $\mathbf{y} = \{y_1, \ldots, y_n\}$ with $y_i \in \{-1, +1\}$.
- Let $\mathbf{f} = \{f_1, \ldots, f_n\}$ be a set of latent variables. Assume that the class labels have a Bernoulli distribution with success probability given by a transformation of the latent variables:

$$p(y_i|f_i) = \Phi(y_i f_i). \tag{1}$$

Here Φ denotes the cumulative function of the Gaussian density; based on this modeling assumption, the likelihood function is:

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i). \tag{2}$$

The latent variables $\mathbf{f}$ are given a zero mean GP prior with covariance K:

$$\mathbf{f} \sim \mathcal{N}(\mathbf{f}|\mathbf{0}, K). \tag{3}$$

- Let $k(\mathbf{x}_i, \mathbf{x}_j|\theta)$ be the function modeling the covariance between latent variables evaluated at the input vectors, parameterized by a vector of hyper-parameters θ.

Fully Bayesian Treatment

- In a fully Bayesian treatment, the aim is to integrate out latent variables as well as hyper-parameters:

$$p(y_*|\mathbf{y}) = \int p(y_*|f_*)p(f_*|\mathbf{f},\theta)p(\mathbf{f},\theta|\mathbf{y})df_*d\mathbf{f}d\theta. \tag{4}$$

Fully Bayesian Treatment

- In a fully Bayesian treatment, the aim is to integrate out latent variables as well as hyper-parameters:

$$p(y_*|\mathbf{y}) = \int p(y_*|f_*)p(f_*|\mathbf{f},\theta)p(\mathbf{f},\theta|\mathbf{y})df_*d\mathbf{f}d\theta. \tag{4}$$

- The integration with respect to f_* can be done analytically, whereas the integration with respect to latent variables and hyper-parameters requires the joint posterior distribution $p(\mathbf{f},\theta|\mathbf{y})$.

Fully Bayesian Treatment

- In a fully Bayesian treatment, the aim is to integrate out latent variables as well as hyper-parameters:

$$p(y_*|\mathbf{y}) = \int p(y_*|f_*)p(f_*|\mathbf{f},\theta)p(\mathbf{f},\theta|\mathbf{y})df_*\,d\mathbf{f}\,d\theta. \tag{4}$$

- The integration with respect to f_* can be done analytically, whereas the integration with respect to latent variables and hyper-parameters requires the joint posterior distribution $p(\mathbf{f},\theta|\mathbf{y})$.
- One way to tackle the intractability in characterizing $p(\mathbf{f},\theta|\mathbf{y})$ is to draw samples from $p(\mathbf{f},\theta|\mathbf{y})$ using MCMC methods, so that a Monte Carlo estimate of the predictive distribution can be used

$$p(y_*|\mathbf{y}) \simeq \frac{1}{N}\sum_{i=1}^{N} \int p(y_*|f_*)p(f_*|\mathbf{f}^{(i)},\theta^{(i)})df_*, \tag{5}$$

where $\mathbf{f}^{(i)},\theta^{(i)}$ denotes the ith sample from $p(\mathbf{f},\theta|\mathbf{y})$.

MCMC Sampling from $p(\mathbf{f}, \theta|\mathbf{y})$

- Typical constructions to draw samples from $p(\mathbf{f}, \theta|\mathbf{y})$, resort to a Gibbs sampler, whereby $\mathbf{f}$ and θ are updated in turn.

MCMC Sampling from $p(\mathbf{f}, \theta|\mathbf{y})$

- Typical constructions to draw samples from $p(\mathbf{f}, \theta|\mathbf{y})$, resort to a Gibbs sampler, whereby $\mathbf{f}$ and θ are updated in turn.
- Drawing samples from $p(\mathbf{f}|\mathbf{y}, \theta)$ achieved via numerous constructions e.g. Elliptical Slice Sampling (ELL-SS)

MCMC Sampling from $p(\mathbf{f}, \theta|\mathbf{y})$

- Typical constructions to draw samples from $p(\mathbf{f}, \theta|\mathbf{y})$, resort to a Gibbs sampler, whereby $\mathbf{f}$ and θ are updated in turn.
- Drawing samples from $p(\mathbf{f}|\mathbf{y}, \theta)$ achieved via numerous constructions e.g. Elliptical Slice Sampling (ELL-SS)
- Simplified version of Riemann manifold Hamiltonian Monte Carlo (RMHMC) which makes it possible to obtain samples from the posterior distribution over $\mathbf{f}$ in $O(n^2)$ once K is factorized.

MCMC Sampling from $p(\mathbf{f}, \theta|\mathbf{y})$

- Typical constructions to draw samples from $p(\mathbf{f}, \theta|\mathbf{y})$, resort to a Gibbs sampler, whereby $\mathbf{f}$ and θ are updated in turn.
- Drawing samples from $p(\mathbf{f}|\mathbf{y}, \theta)$ achieved via numerous constructions e.g. Elliptical Slice Sampling (ELL-SS)
- Simplified version of Riemann manifold Hamiltonian Monte Carlo (RMHMC) which makes it possible to obtain samples from the posterior distribution over $\mathbf{f}$ in $O(n^2)$ once K is factorized.
- Drawing samples from the posterior over θ notoriously challenging due to coupling of latent variable and covariance parameters

MCMC Sampling from $p(\mathbf{f}, \theta|\mathbf{y})$

- Typical constructions to draw samples from $p(\mathbf{f}, \theta|\mathbf{y})$, resort to a Gibbs sampler, whereby $\mathbf{f}$ and θ are updated in turn.
- Drawing samples from $p(\mathbf{f}|\mathbf{y}, \theta)$ achieved via numerous constructions e.g. Elliptical Slice Sampling (ELL-SS)
- Simplified version of Riemann manifold Hamiltonian Monte Carlo (RMHMC) which makes it possible to obtain samples from the posterior distribution over $\mathbf{f}$ in $O(n^2)$ once K is factorized.
- Drawing samples from the posterior over θ notoriously challenging due to coupling of latent variable and covariance parameters
- Reparametrisation to reduce effect of coupling - Centered, Non-Centered, Sufficent and Auxiliary Augmentation, Surrogate Data model

MCMC Sampling from $p(\mathbf{f}, \theta|\mathbf{y})$

- Typical constructions to draw samples from $p(\mathbf{f}, \theta|\mathbf{y})$, resort to a Gibbs sampler, whereby $\mathbf{f}$ and θ are updated in turn.
- Drawing samples from $p(\mathbf{f}|\mathbf{y}, \theta)$ achieved via numerous constructions e.g. Elliptical Slice Sampling (ELL-SS)
- Simplified version of Riemann manifold Hamiltonian Monte Carlo (RMHMC) which makes it possible to obtain samples from the posterior distribution over $\mathbf{f}$ in $O(n^2)$ once K is factorized.
- Drawing samples from the posterior over θ notoriously challenging due to coupling of latent variable and covariance parameters
- Reparametrisation to reduce effect of coupling - Centered, Non-Centered, Sufficent and Auxiliary Augmentation, Surrogate Data model
- Intuitively, the best strategy to break the correlation between latent variables and hyper-parameters would be to integrate out the latent variables altogether.

MCMC Sampling from $p(\mathbf{f}, \theta|\mathbf{y})$

- Typical constructions to draw samples from $p(\mathbf{f}, \theta|\mathbf{y})$, resort to a Gibbs sampler, whereby $\mathbf{f}$ and θ are updated in turn.
- Drawing samples from $p(\mathbf{f}|\mathbf{y}, \theta)$ achieved via numerous constructions e.g. Elliptical Slice Sampling (ELL-SS)
- Simplified version of Riemann manifold Hamiltonian Monte Carlo (RMHMC) which makes it possible to obtain samples from the posterior distribution over $\mathbf{f}$ in $O(n^2)$ once K is factorized.
- Drawing samples from the posterior over θ notoriously challenging due to coupling of latent variable and covariance parameters
- Reparametrisation to reduce effect of coupling - Centered, Non-Centered, Sufficent and Auxiliary Augmentation, Surrogate Data model
- Intuitively, the best strategy to break the correlation between latent variables and hyper-parameters would be to integrate out the latent variables altogether.
- This is not possible, but a strategy is presented that uses an unbiased estimate of the marginal likelihood $p(\mathbf{y}|\theta)$ to devise an MCMC strategy producing samples from the correct posterior distribution $p(\theta|\mathbf{y})$.

MCMC Sampling from $p(\mathbf{f}, \theta|\mathbf{y})$

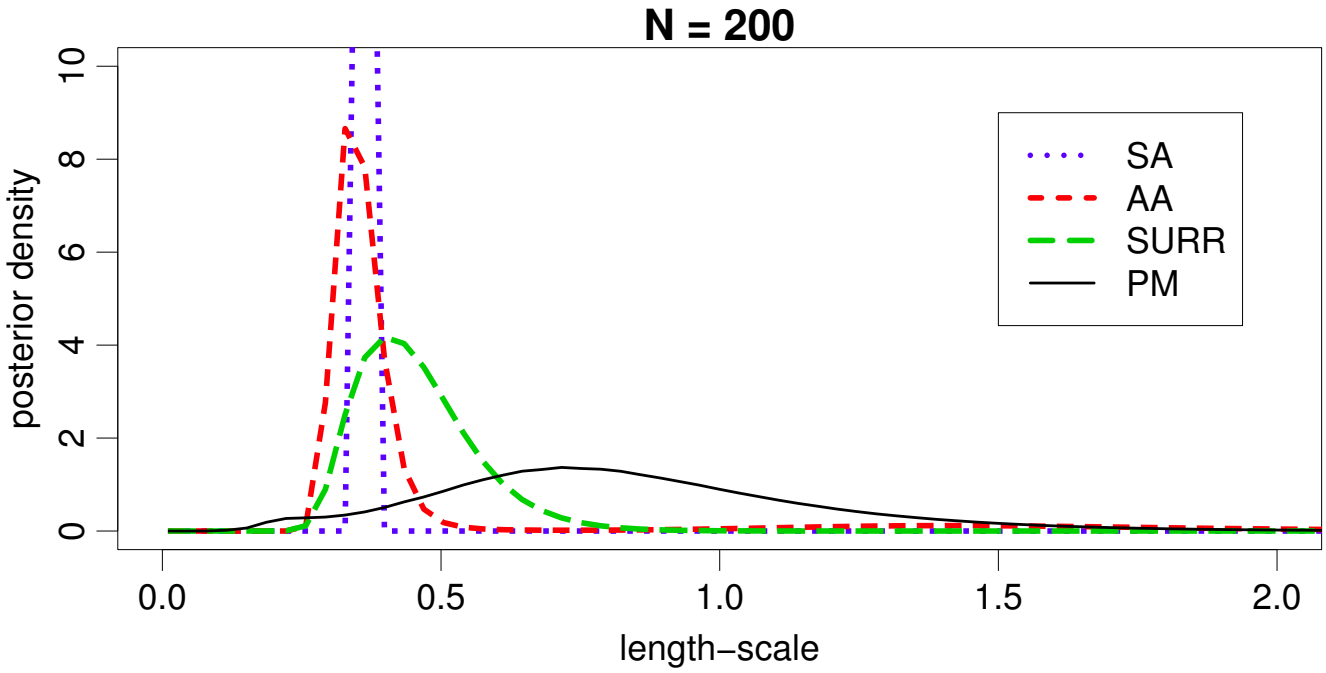


Figure : Comparison of the posterior distribution $p(\theta|\mathbf{y})$ with the posterior $p(\theta|\mathbf{f})$ in the SA parameterization, the posterior $p(\theta|\mathbf{y}, \nu)$ in the AA parameterization, and the parameterization used in the SURR method.

MCMC Sampling from $p(\mathbf{f}, \theta|\mathbf{y})$ the Pseudo-Marginal Approach

- We are interested in sampling from the posterior distribution

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta). \tag{6}$$

MCMC Sampling from $p(\mathbf{f}, \theta|\mathbf{y})$ the Pseudo-Marginal Approach

- We are interested in sampling from the posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{6}$$

- In order to do that, we would need to integrate out the latent variables:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f} \tag{7}$$

  and use this along with the prior $p(\boldsymbol{\theta})$ in the Hastings ratio:

$$z = \frac{p(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \frac{\pi(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})} \tag{8}$$

MCMC Sampling from $p(\mathbf{f}, \theta|\mathbf{y})$ the Pseudo-Marginal Approach

- We are interested in sampling from the posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{6}$$

- In order to do that, we would need to integrate out the latent variables:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f} \tag{7}$$

  and use this along with the prior $p(\boldsymbol{\theta})$ in the Hastings ratio:

$$z = \frac{p(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \frac{\pi(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})} \tag{8}$$

- As already discussed, analytically integrating out $\mathbf{f}$ is not possible.

MCMC Sampling from $p(\mathbf{f}, \theta|\mathbf{y})$ the Pseudo-Marginal Approach

- We are interested in sampling from the posterior distribution

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta). \tag{6}$$

- In order to do that, we would need to integrate out the latent variables:

$$p(\mathbf{y}|\theta) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\theta)d\mathbf{f} \tag{7}$$

  and use this along with the prior $p(\theta)$ in the Hastings ratio:

$$z = \frac{p(\mathbf{y}|\theta')p(\theta')}{p(\mathbf{y}|\theta)p(\theta)} \frac{\pi(\theta|\theta')}{\pi(\theta'|\theta)} \tag{8}$$

- As already discussed, analytically integrating out $\mathbf{f}$ is not possible.
- The results in Andrieu and Roberts, 2009 show that we can plug into the Hastings ratio an estimate $\tilde{p}(\mathbf{y}|\theta)$ of the marginal $p(\mathbf{y}|\theta)$

MCMC Sampling from $p(\mathbf{f}, \theta|\mathbf{y})$ the Pseudo-Marginal Approach

- We are interested in sampling from the posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{6}$$

- In order to do that, we would need to integrate out the latent variables:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f} \tag{7}$$

  and use this along with the prior $p(\boldsymbol{\theta})$ in the Hastings ratio:

$$z = \frac{p(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \frac{\pi(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})} \tag{8}$$

- As already discussed, analytically integrating out $\mathbf{f}$ is not possible.
- The results in Andrieu and Roberts, 2009 show that we can plug into the Hastings ratio an estimate $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ of the marginal $p(\mathbf{y}|\boldsymbol{\theta})$
- As long as this is positive and unbiased, then the sampler will draw samples from the correct posterior $p(\boldsymbol{\theta}|\mathbf{y})$.

$$\tilde{z} = \frac{\tilde{p}(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{\tilde{p}(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \frac{\pi(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})} \tag{9}$$

Unbiased estimation of $p(\mathbf{y}|\theta)$ using importance sampling

- In order to obtain an unbiased estimator $\tilde{p}(\mathbf{y}|\theta)$ for the marginal $p(\mathbf{y}|\theta)$, employ importance sampling.

Unbiased estimation of $p(\mathbf{y}|\theta)$ using importance sampling

- In order to obtain an unbiased estimator $\tilde{p}(\mathbf{y}|\theta)$ for the marginal $p(\mathbf{y}|\theta)$, employ importance sampling.
- Draw N_{imp} samples $\mathbf{f}_i$ from the approximating distribution $q(\mathbf{f}|\mathbf{y}, \theta)$, so to approximate the marginal $p(\mathbf{y}|\theta)$ by:

Unbiased estimation of $p(\mathbf{y}|\theta)$ using importance sampling

- In order to obtain an unbiased estimator $\tilde{p}(\mathbf{y}|\theta)$ for the marginal $p(\mathbf{y}|\theta)$, employ importance sampling.
- Draw N_{imp} samples $\mathbf{f}_i$ from the approximating distribution $q(\mathbf{f}|\mathbf{y}, \theta)$, so to approximate the marginal $p(\mathbf{y}|\theta)$ by:

$$\tilde{p}(\mathbf{y}|\theta) \simeq \frac{1}{N_{\text{imp}}} \sum_{i=1}^{N_{\text{imp}}} \frac{p(\mathbf{y}|\mathbf{f}_i)p(\mathbf{f}_i|\theta)}{q(\mathbf{f}_i|\mathbf{y}, \theta)} \tag{10}$$

Unbiased estimation of $p(\mathbf{y}|\theta)$ using importance sampling

- In order to obtain an unbiased estimator $\tilde{p}(\mathbf{y}|\theta)$ for the marginal $p(\mathbf{y}|\theta)$, employ importance sampling.
- Draw N_{imp} samples $\mathbf{f}_i$ from the approximating distribution $q(\mathbf{f}|\mathbf{y}, \theta)$, so to approximate the marginal $p(\mathbf{y}|\theta)$ by:

$$\tilde{p}(\mathbf{y}|\theta) \simeq \frac{1}{N_{\text{imp}}} \sum_{i=1}^{N_{\text{imp}}} \frac{p(\mathbf{y}|\mathbf{f}_i)p(\mathbf{f}_i|\theta)}{q(\mathbf{f}_i|\mathbf{y}, \theta)} \tag{10}$$

- Exploit approximate posterior constructions such as Laplace, Variational, Expectation Propagation for $q(\mathbf{f}_i|\mathbf{y}, \theta)$

Unbiased estimation of $p(\mathbf{y}|\theta)$ using importance sampling

- In order to obtain an unbiased estimator $\tilde{p}(\mathbf{y}|\theta)$ for the marginal $p(\mathbf{y}|\theta)$, employ importance sampling.
- Draw N_{imp} samples $\mathbf{f}_i$ from the approximating distribution $q(\mathbf{f}|\mathbf{y}, \theta)$, so to approximate the marginal $p(\mathbf{y}|\theta)$ by:

$$\tilde{p}(\mathbf{y}|\theta) \simeq \frac{1}{N_{\text{imp}}} \sum_{i=1}^{N_{\text{imp}}} \frac{p(\mathbf{y}|\mathbf{f}_i)p(\mathbf{f}_i|\theta)}{q(\mathbf{f}_i|\mathbf{y}, \theta)} \tag{10}$$

- Exploit approximate posterior constructions such as Laplace, Variational, Expectation Propagation for $q(\mathbf{f}_i|\mathbf{y}, \theta)$
- Filiponne and Girolami (IEEE trans PAMI, 2014) evaluate a number of approximating distributions

Unbiased estimation of $p(\mathbf{y}|\theta)$ using importance sampling

- In order to obtain an unbiased estimator $\tilde{p}(\mathbf{y}|\theta)$ for the marginal $p(\mathbf{y}|\theta)$, employ importance sampling.
- Draw N_{imp} samples $\mathbf{f}_i$ from the approximating distribution $q(\mathbf{f}|\mathbf{y},\theta)$, so to approximate the marginal $p(\mathbf{y}|\theta)$ by:

$$\tilde{p}(\mathbf{y}|\theta) \simeq \frac{1}{N_{\text{imp}}} \sum_{i=1}^{N_{\text{imp}}} \frac{p(\mathbf{y}|\mathbf{f}_i)p(\mathbf{f}_i|\theta)}{q(\mathbf{f}_i|\mathbf{y},\theta)} \tag{10}$$

- Exploit approximate posterior constructions such as Laplace, Variational, Expectation Propagation for $q(\mathbf{f}_i|\mathbf{y},\theta)$
- Filiponne and Girolami (IEEE trans PAMI, 2014) evaluate a number of approximating distributions
- Expectation Propagation consistently superior in controlling estimator variance

Unbiased estimation of $p(\mathbf{y}|\theta)$ using importance sampling

- In order to obtain an unbiased estimator $\tilde{p}(\mathbf{y}|\theta)$ for the marginal $p(\mathbf{y}|\theta)$, employ importance sampling.
- Draw N_{imp} samples $\mathbf{f}_i$ from the approximating distribution $q(\mathbf{f}|\mathbf{y},\theta)$, so to approximate the marginal $p(\mathbf{y}|\theta)$ by:

$$\tilde{p}(\mathbf{y}|\theta) \simeq \frac{1}{N_{\text{imp}}} \sum_{i=1}^{N_{\text{imp}}} \frac{p(\mathbf{y}|\mathbf{f}_i)p(\mathbf{f}_i|\theta)}{q(\mathbf{f}_i|\mathbf{y},\theta)} \tag{10}$$

- Exploit approximate posterior constructions such as Laplace, Variational, Expectation Propagation for $q(\mathbf{f}_i|\mathbf{y},\theta)$
- Filiponne and Girolami (IEEE trans PAMI, 2014) evaluate a number of approximating distributions
- Expectation Propagation consistently superior in controlling estimator variance unsurprising from empirical evidence in other applications

Unbiased estimation of $p(\mathbf{y}|\theta)$ using importance sampling

- In order to obtain an unbiased estimator $\tilde{p}(\mathbf{y}|\theta)$ for the marginal $p(\mathbf{y}|\theta)$, employ importance sampling.
- Draw N_{imp} samples $\mathbf{f}_i$ from the approximating distribution $q(\mathbf{f}|\mathbf{y}, \theta)$, so to approximate the marginal $p(\mathbf{y}|\theta)$ by:

$$\tilde{p}(\mathbf{y}|\theta) \simeq \frac{1}{N_{\text{imp}}} \sum_{i=1}^{N_{\text{imp}}} \frac{p(\mathbf{y}|\mathbf{f}_i)p(\mathbf{f}_i|\theta)}{q(\mathbf{f}_i|\mathbf{y}, \theta)} \tag{10}$$

- Exploit approximate posterior constructions such as Laplace, Variational, Expectation Propagation for $q(\mathbf{f}_i|\mathbf{y}, \theta)$
- Filiponne and Girolami (IEEE trans PAMI, 2014) evaluate a number of approximating distributions
- Expectation Propagation consistently superior in controlling estimator variance unsurprising from empirical evidence in other applications but lacks analytic support

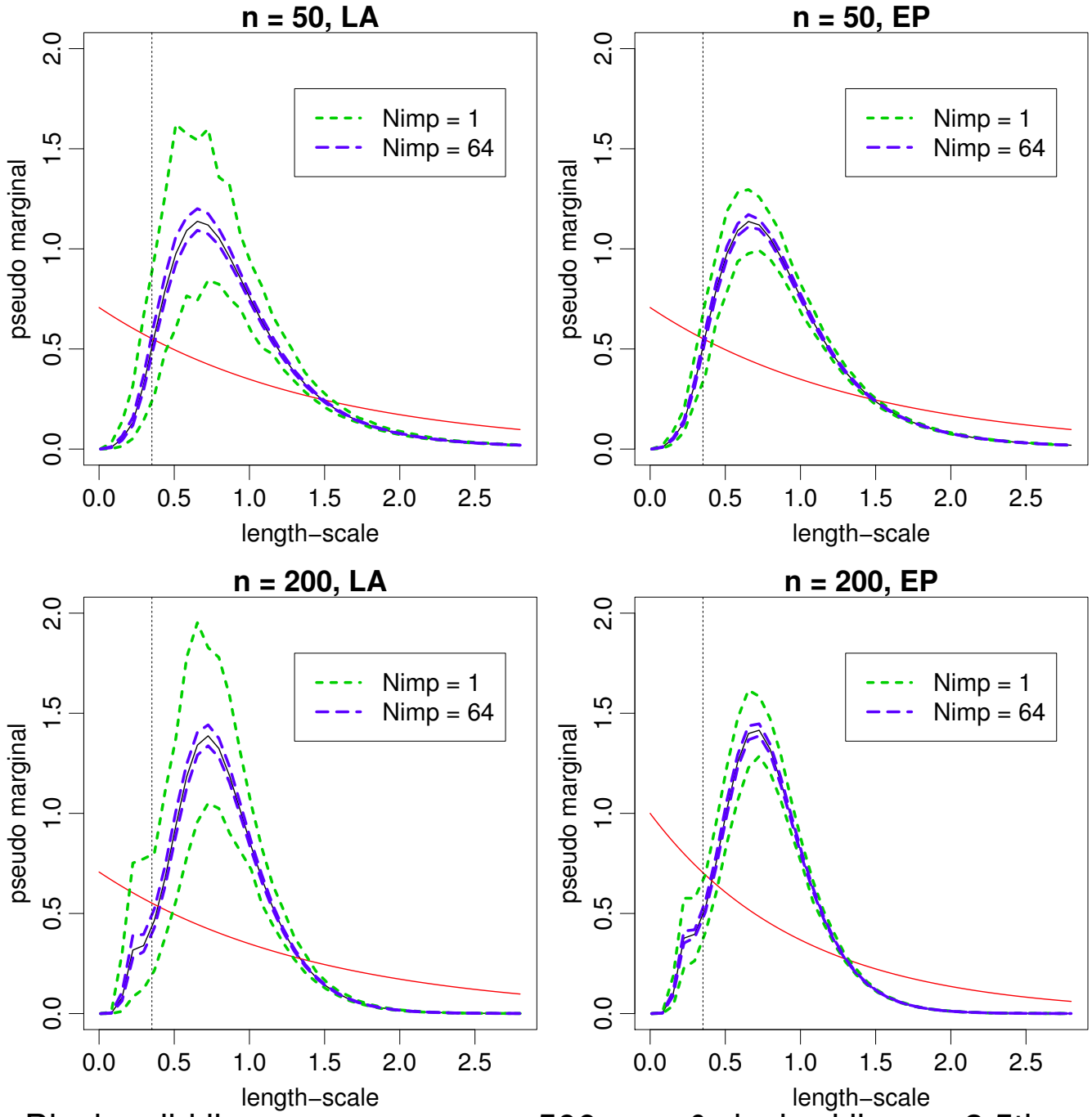


Figure : Black solid lines = average over 500 reps & dashed lines = 2.5th and 97.5th quantiles for $N_{\text{imp}} = 1$ and $N_{\text{imp}} = 64$. The solid red line is the prior density.

Abalone $n = 2835$

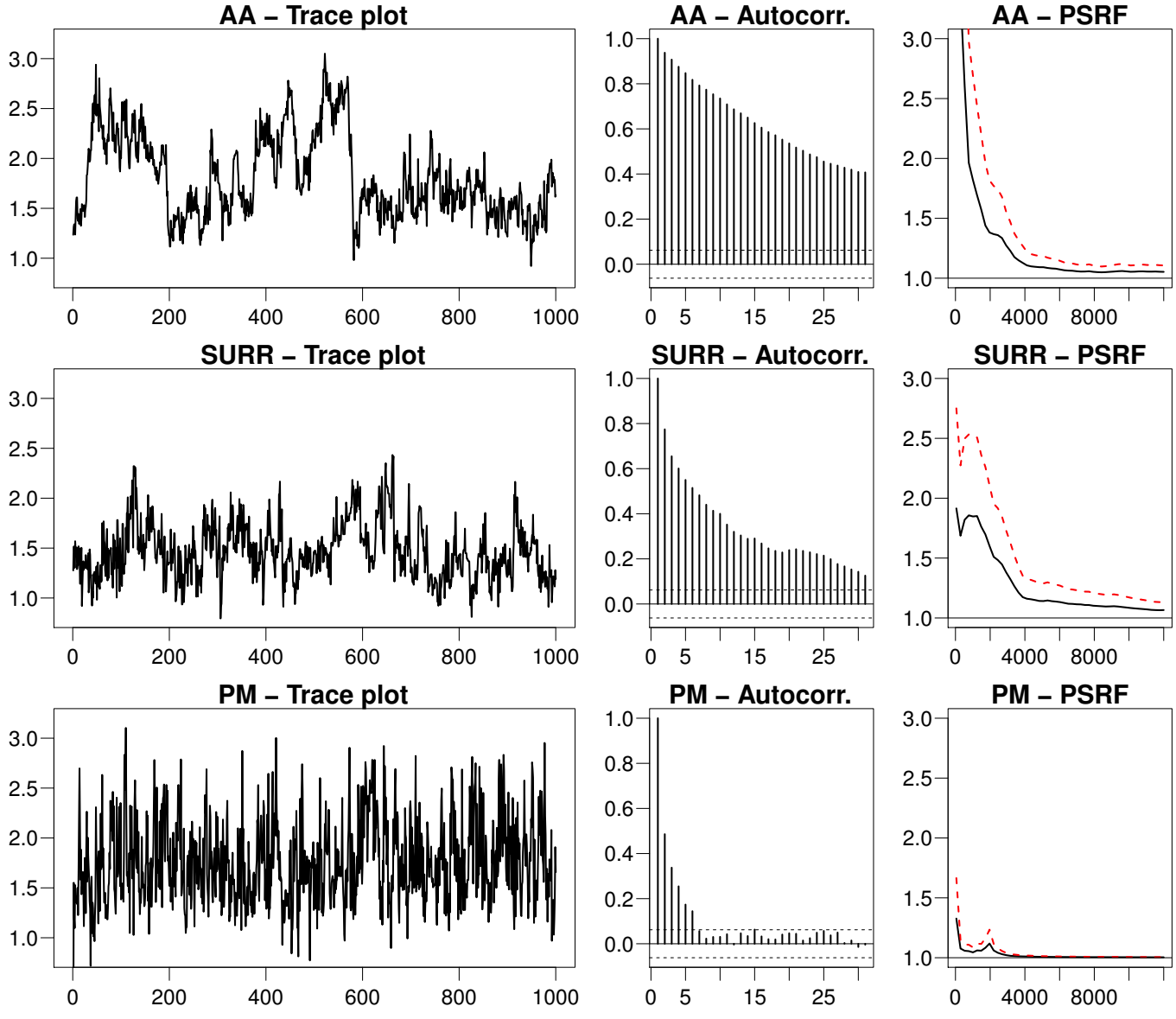


Figure : Summary of efficiency and convergence speed on the Abalone data set.

Breast $n = 682$

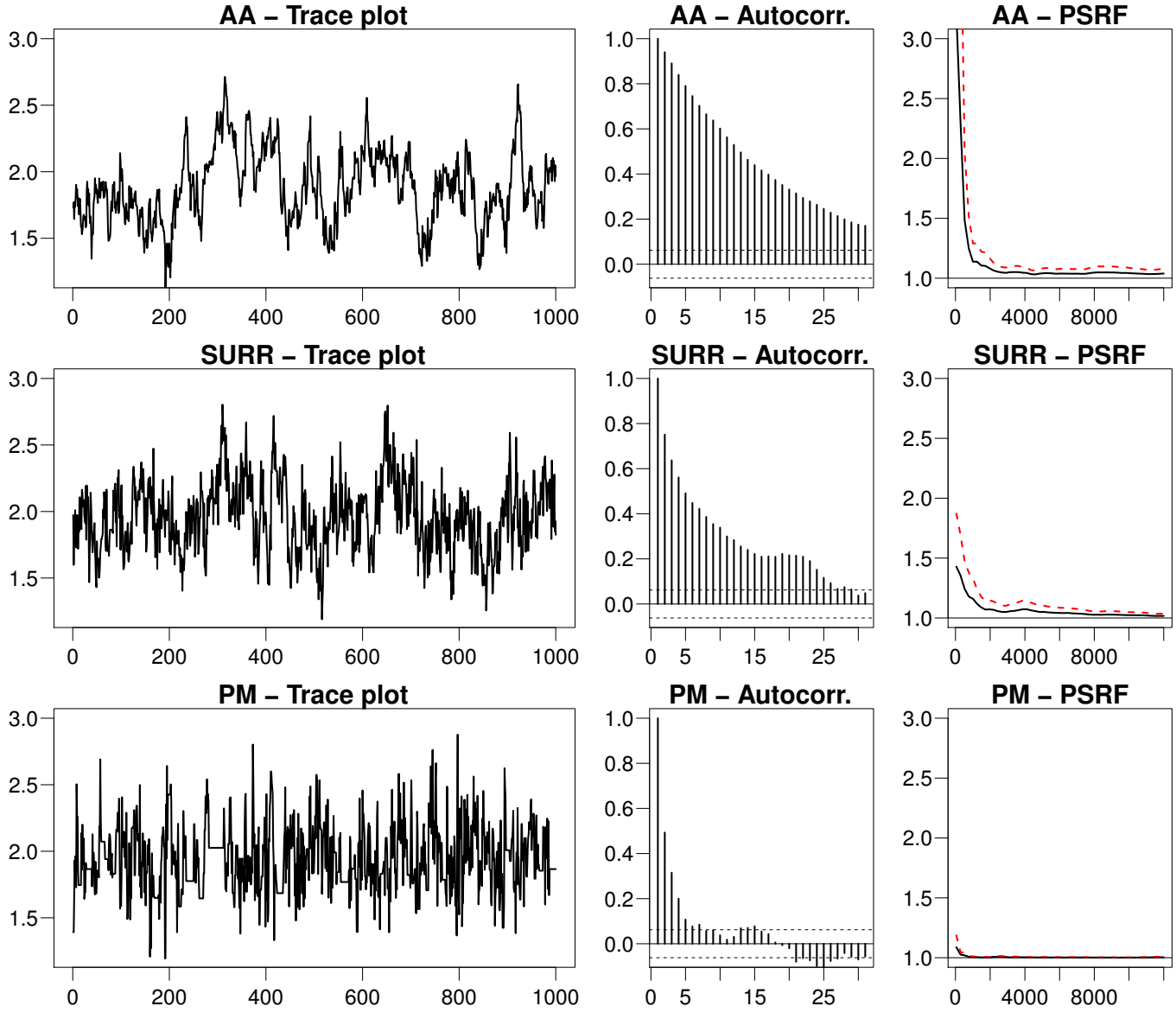


Figure : Summary of efficiency and convergence speed on the Breast data set.

Pima $n = 768$

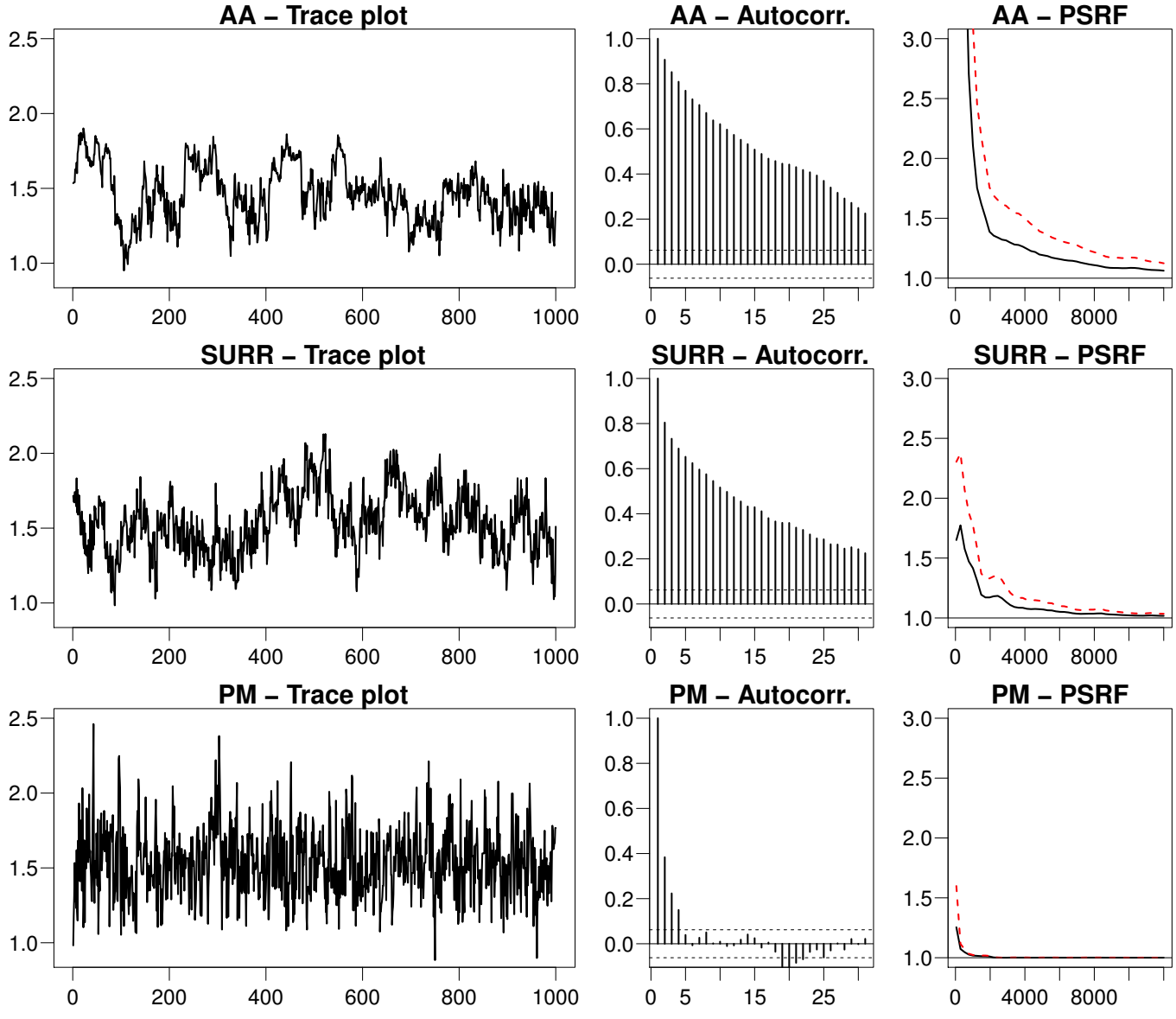
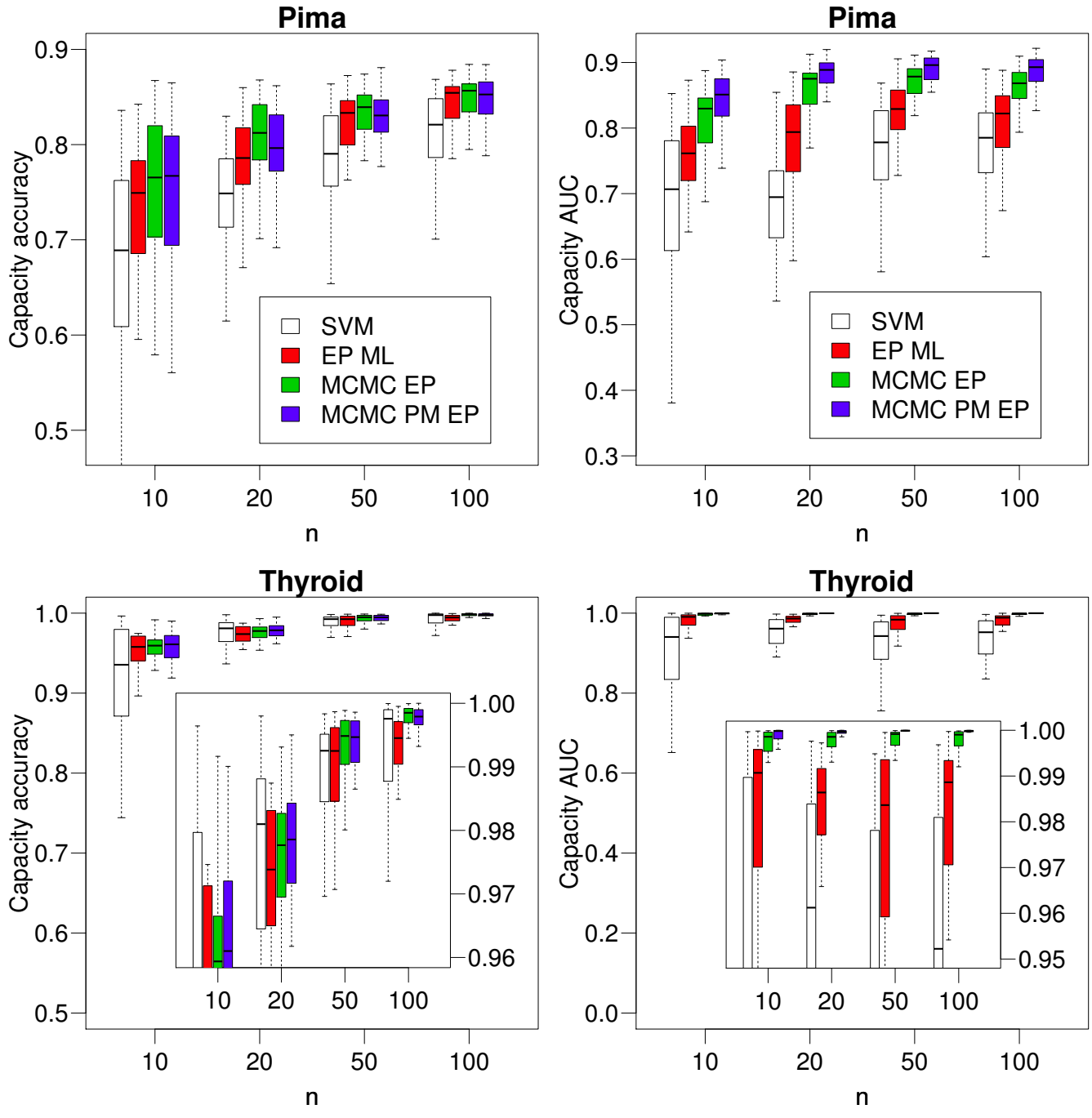
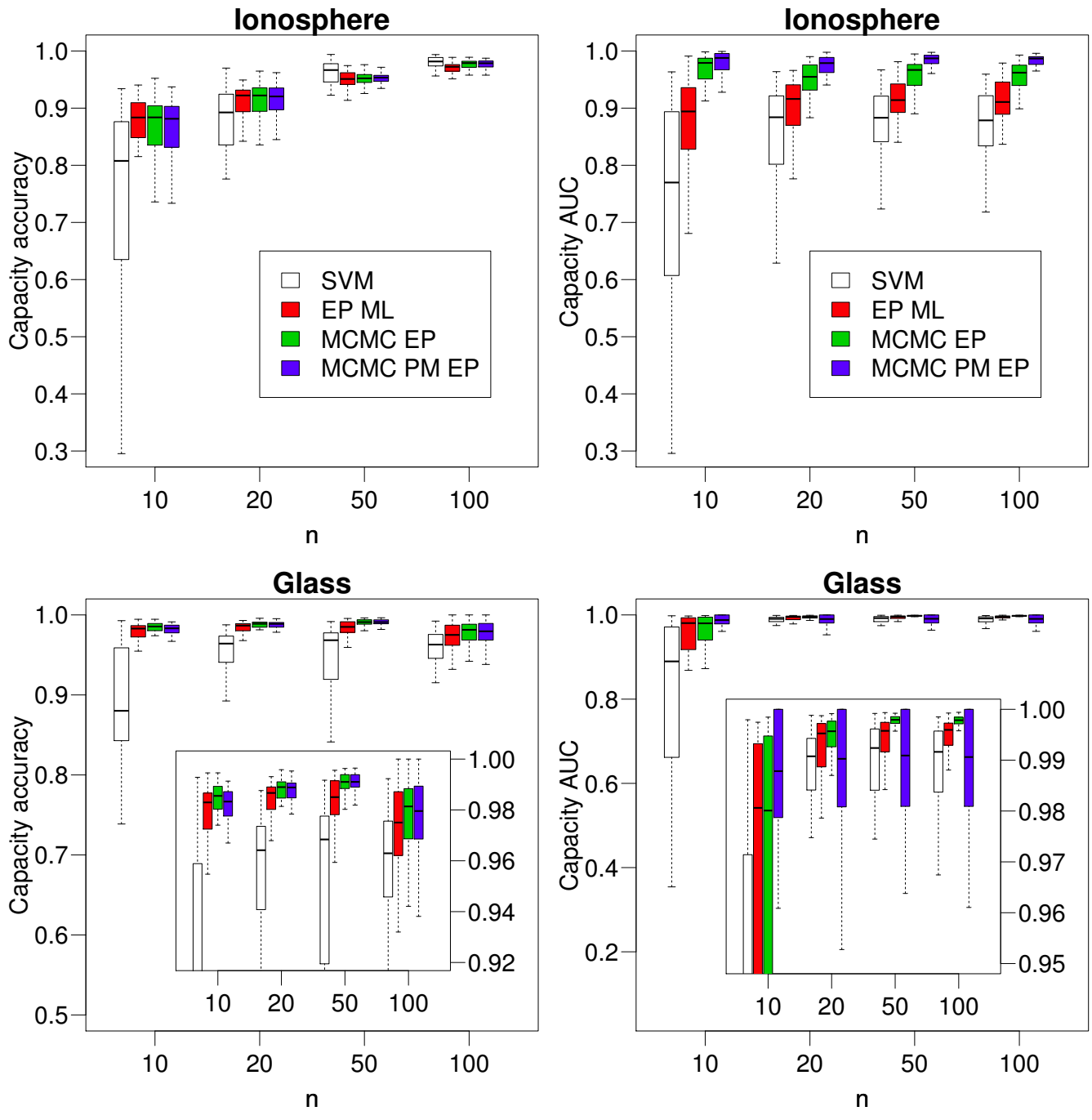


Figure : Summary of efficiency and convergence speed on the Pima data set.


Pima
Capacity accuracy
0.9
0.8
0.7
0.6
0.5
SVM
EP ML
MCMC EP
MCMC PM EP
10
20
50
100
n
Pima
Capacity AUC
0.9
0.8
0.7
0.6
0.5
0.4
0.3
SVM
EP ML
MCMC EP
MCMC PM EP
10
20
50
100
n
Thyroid
Capacity accuracy
1.0
0.9
0.8
0.7
0.6
0.5
1.00
0.99
0.98
0.97
0.96
10
20
50
100
10
20
50
100
n
Thyroid
Capacity AUC
1.0
0.8
0.6
0.4
0.2
0.0
1.00
0.99
0.98
0.97
0.96
0.95
10
20
50
100
10
20
50
100
n


Ionosphere
Capacity accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
10
20
50
100
n
SVM
EP ML
MCMC EP
MCMC PM EP
Ionosphere
Capacity AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
10
20
50
100
n
SVM
EP ML
MCMC EP
MCMC PM EP
Glass
Capacity accuracy
1.0
0.9
0.8
0.7
0.6
0.5
1.00
0.98
0.96
0.94
0.92
10
20
50
100
10
20
50
100
n
Glass
Capacity AUC
1.0
0.8
0.6
0.4
0.2
1.00
0.99
0.98
0.97
0.96
0.95
10
20
50
100
10
20
50
100
n

Integrating Different Image Regions

- Distinct feature representation of X, $\mathcal{F}_j(X) = \mathbf{x}(j)$, is nonlinearly transformed such that $f_j(\mathbf{x}(j)) : \mathcal{F}_j \mapsto \mathbb{R}$.

Integrating Different Image Regions

- Distinct feature representation of X, $\mathcal{F}_j(X) = \mathbf{x}(j)$, is nonlinearly transformed such that $f_j(\mathbf{x}(j)) : \mathcal{F}_j \mapsto \mathbb{R}$.
- A linear model is employed in this new space such that the overall nonlinear transformation is

$$f(X) = \sum_{j=1}^{\mathcal{J}} \beta_j f_j(\mathbf{x}(j))$$

Integrating Different Image Regions

- Distinct feature representation of X, $\mathcal{F}_j(X) = \mathbf{x}(j)$, is nonlinearly transformed such that $f_j(\mathbf{x}(j)) : \mathcal{F}_j \mapsto \mathbb{R}$.
- A linear model is employed in this new space such that the overall nonlinear transformation is

$$f(X) = \sum_{j=1}^{\mathcal{J}} \beta_j f_j(\mathbf{x}(j))$$

- Where each $f_j(\mathbf{x}(j)) \sim GP(\boldsymbol{\theta}_j)$ where $GP(\boldsymbol{\theta}_j)$ corresponds to a Gaussian process with mean and covariance functions $m_j(\mathbf{x}(j))$ and $C_j(\mathbf{x}(j), \mathbf{x}'(j); \boldsymbol{\theta}_j)$

Integrating Different Image Regions

- Distinct feature representation of X, $\mathcal{F}_j(X) = \mathbf{x}(j)$, is nonlinearly transformed such that $f_j(\mathbf{x}(j)) : \mathcal{F}_j \mapsto \mathbb{R}$.
- A linear model is employed in this new space such that the overall nonlinear transformation is

$$f(X) = \sum_{j=1}^{\mathcal{J}} \beta_j f_j(\mathbf{x}(j))$$

- Where each $f_j(\mathbf{x}(j)) \sim GP(\boldsymbol{\theta}_j)$ where $GP(\boldsymbol{\theta}_j)$ corresponds to a Gaussian process with mean and covariance functions $m_j(\mathbf{x}(j))$ and $C_j(\mathbf{x}(j), \mathbf{x}'(j); \boldsymbol{\theta}_j)$
- Then $f(X) \sim GP\left(\boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_{\mathcal{J}}, \beta_1 \cdots \beta_{\mathcal{J}}\right)$ where now the overall mean and covariance functions follow as

$$\sum_{j=1}^{\mathcal{J}} \beta_j m_j(\mathbf{x}(j)) \quad \sum_{j=1}^{\mathcal{J}} \beta_j^2 C_j(\mathbf{x}(j), \mathbf{x}'(j); \boldsymbol{\theta}_j)$$

Integrating Different Image Regions

- Distinct feature representation of X, $\mathcal{F}_j(X) = \mathbf{x}(j)$, is nonlinearly transformed such that $f_j(\mathbf{x}(j)) : \mathcal{F}_j \mapsto \mathbb{R}$.
- A linear model is employed in this new space such that the overall nonlinear transformation is

$$f(X) = \sum_{j=1}^{\mathcal{J}} \beta_j f_j(\mathbf{x}(j))$$

- Where each $f_j(\mathbf{x}(j)) \sim GP(\boldsymbol{\theta}_j)$ where $GP(\boldsymbol{\theta}_j)$ corresponds to a Gaussian process with mean and covariance functions $m_j(\mathbf{x}(j))$ and $C_j(\mathbf{x}(j), \mathbf{x}'(j); \boldsymbol{\theta}_j)$
- Then $f(X) \sim GP\left(\boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_{\mathcal{J}}, \beta_1 \cdots \beta_{\mathcal{J}}\right)$ where now the overall mean and covariance functions follow as

$$\sum_{j=1}^{\mathcal{J}} \beta_j m_j(\mathbf{x}(j)) \quad \sum_{j=1}^{\mathcal{J}} \beta_j^2 C_j(\mathbf{x}(j), \mathbf{x}'(j); \boldsymbol{\theta}_j)$$

- Posterior over $\beta_1 \cdots \beta_{\mathcal{J}}$ suggestive of relative importance in terms of predictive power of each distinct region

Integrating Different Image Regions

- All disease groups could be discriminated simultaneously with high accuracy using the subcortical motor network.

Integrating Different Image Regions

- All disease groups could be discriminated simultaneously with high accuracy using the subcortical motor network.
- The region providing the most accurate predictions overall was the midbrain/brainstem, which discriminated all disease groups from one another and from HCs.

Integrating Different Image Regions

- All disease groups could be discriminated simultaneously with high accuracy using the subcortical motor network.
- The region providing the most accurate predictions overall was the midbrain/brainstem, which discriminated all disease groups from one another and from HCs.
- The subcortical network also produced more accurate predictions than the whole brain and all of its constituent regions.

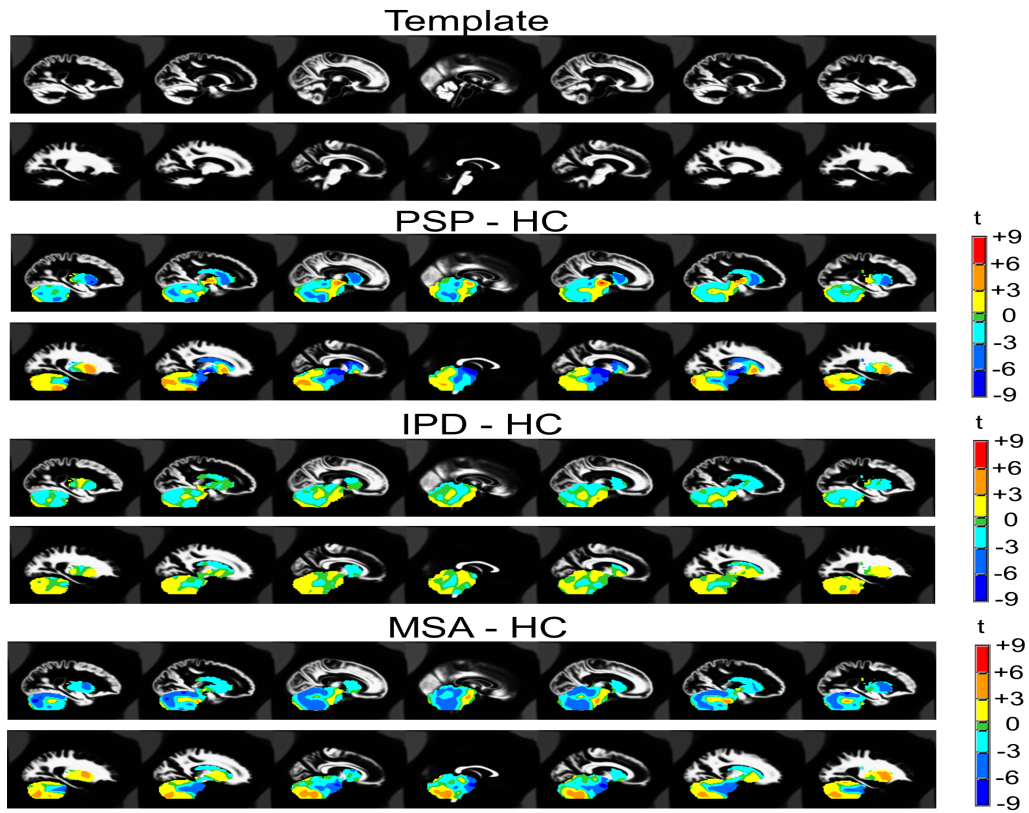
Integrating Different Image Regions

- All disease groups could be discriminated simultaneously with high accuracy using the subcortical motor network.
- The region providing the most accurate predictions overall was the midbrain/brainstem, which discriminated all disease groups from one another and from HCs.
- The subcortical network also produced more accurate predictions than the whole brain and all of its constituent regions.
- PSP was accurately predicted from the midbrain/brainstem, cerebellum and all basal ganglia compartments; MSA from the midbrain/brainstem and cerebellum and IPD from the midbrain/brainstem only.

Integrating Different Image Regions

- All disease groups could be discriminated simultaneously with high accuracy using the subcortical motor network.
- The region providing the most accurate predictions overall was the midbrain/brainstem, which discriminated all disease groups from one another and from HCs.
- The subcortical network also produced more accurate predictions than the whole brain and all of its constituent regions.
- PSP was accurately predicted from the midbrain/brainstem, cerebellum and all basal ganglia compartments; MSA from the midbrain/brainstem and cerebellum and IPD from the midbrain/brainstem only.
- This study demonstrates that automated analysis of structural MRI can accurately predict diagnosis in individual patients with Parkinsonian disorders, and identifies distinct patterns of regional atrophy particularly useful for this process.

Posterior Mean Region Weights



Discussion and Outlook

- Small scale preliminary proof-of-concept study

Discussion and Outlook

- Small scale preliminary proof-of-concept study
- Clinical importance high and ongoing collaboration with more extensive cohort studies

Discussion and Outlook

- Small scale preliminary proof-of-concept study
- Clinical importance high and ongoing collaboration with more extensive cohort studies
- Study exploits hierarchic model enabling integration of heterogeneous data sources

Discussion and Outlook

- Small scale preliminary proof-of-concept study
- Clinical importance high and ongoing collaboration with more extensive cohort studies
- Study exploits hierarchic model enabling integration of heterogeneous data sources
- Gets around the poor mixing problem in hierarchic Bayesian models targeting marginal posterior directly

Discussion and Outlook

- Small scale preliminary proof-of-concept study
- Clinical importance high and ongoing collaboration with more extensive cohort studies
- Study exploits hierarchic model enabling integration of heterogeneous data sources
- Gets around the poor mixing problem in hierarchic Bayesian models targeting marginal posterior directly
- Pseudo-marginal construction exploits approximate posteriors in importance sampling

Discussion and Outlook

- Small scale preliminary proof-of-concept study
- Clinical importance high and ongoing collaboration with more extensive cohort studies
- Study exploits hierarchic model enabling integration of heterogeneous data sources
- Gets around the poor mixing problem in hierarchic Bayesian models targeting marginal posterior directly
- Pseudo-marginal construction exploits approximate posteriors in importance sampling
- For GP prior based models PM approach more effective than transformation methods

Discussion and Outlook

- Small scale preliminary proof-of-concept study
- Clinical importance high and ongoing collaboration with more extensive cohort studies
- Study exploits hierarchic model enabling integration of heterogeneous data sources
- Gets around the poor mixing problem in hierarchic Bayesian models targeting marginal posterior directly
- Pseudo-marginal construction exploits approximate posteriors in importance sampling
- For GP prior based models PM approach more effective than transformation methods
- Wider application obvious.

Publications

- Filippone, M. and Girolami, M. Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

Publications

- Filippone, M. and Girolami, M. Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.
- Filippone, M.; Marquand, Andre; Blain, C. R. V.; Williams, S. C. R.; Mourao-Miranda, J.; Girolami, M. Probabilistic Prediction of Neurological Disorders with a Statistical Assessment of Neuroimaging Data Modalities. *Annals of Applied Statistics*, Vol. 6, No. 4, p. 1883-1905, 2012.

Publications

- Filippone, M. and Girolami, M. Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.
- Filippone, M.; Marquand, Andre; Blain, C. R. V.; Williams, S. C. R.; Mourao-Miranda, J.; Girolami, M. Probabilistic Prediction of Neurological Disorders with a Statistical Assessment of Neuroimaging Data Modalities. *Annals of Applied Statistics*, Vol. 6, No. 4, p. 1883-1905, 2012.
- Girolami,M. and Zhong,M. Data Integration for Classification Problems Employing Gaussian Process Priors. Twentieth Annual Conference on Neural Information Processing Systems, *NIPS 19*, (MIT Press), 465 - 472, 2007.