Dense Message Passing for Sparse Principal Component Analysis

Kevin Sharp   Magnus Rattray

University of Manchester

SuSTaIn Sparsity Workshop, Bristol, 16th June 2010
Outline

1. Introduction
   - Motivating Application: Gene Regulation

2. Dense Message Passing
   - Model Description
   - Algorithm Description
   - Statistical Mechanics Theory

3. Results
   - Simulated Data
   - Gene Expression Data
   - Marginal Likelihood Estimation

4. Summary
Gene regulation - inference of explanatory factors.

- Microarray data - ‘Large $p$ small $N$’ regime.
- Explanatory factors have truly sparse loadings.
- Zero-norm priors allocate probability mass to truly sparse solutions.
- Easy to encode prior knowledge of sparse structure.
- But zero-norm priors are problematic for inference.
For the $n^{th}$ data point, $y_n$, we assume:

$$y_n = w x_n + \epsilon_n,$$

where $x_n \sim \mathcal{N}(0, 1)$.

To simplify the description, $\epsilon_n \sim \mathcal{N}(0, I)$.

Integrate out $x$:

$$P(y_n | w) = \mathcal{N}(y_n | 0, I + ww^T)$$
For the $n^{th}$ data point, $y_n$, we assume:

$$y_n = w x_n + \epsilon_n ,$$

where $x_n \sim \mathcal{N}(0, 1)$.

To simplify the description, $\epsilon_n \sim \mathcal{N}(0, I)$.

Integrate out $x$:

$$P(y_n \mid w) = \mathcal{N}(y_n \mid 0, I + w w^T)$$
We use a **spike and slab** mixture prior:

\[ P(w|C, \lambda) = \prod_{j=1}^{p} \left[ (1 - C)\delta(w_j) + C\mathcal{N}(w_j|0, (\lambda)^{-1}) \right] \]

- **C** - Fraction of non-zero \( w \)s
- **\lambda** - inverse width
Express in factorised form using binary variables $z_j \in \{0, 1\}$:

$$P(v, z) = \prod_{j=1}^{p} \left\{ (1 - C) \mathcal{N}(v_j | 0, 1) \right\}^{1-z_j} \left\{ C \mathcal{N}(v_j | 0, \lambda^{-1}) \right\}^{z_j}$$

where $w_j = z_j v_j$ and $v_j \in \mathcal{R}$
Form of the Prior in High Dimensions

- \( z_j \sim \text{Bernoulli} (C) \)

- \( \sum_j z_j \sim \text{Bin}(p, C) \)

For large dimension, \( p \), the fraction of non-zero parameters is highly peaked at \( C \).
Form of the Prior in High Dimensions

- $P(w_j | z_j = 1) = \mathcal{N}(w_j | 0, \lambda^{-1})$

- $w | z \sim \mathcal{N}(0, \lambda^{-1} I)$

- For large dimension, $p$, this distribution is approximately spherical with radius $\sqrt{Cp/\lambda}$. 
- **Conclusion** - A **constraint-based** prior:

\[
p(w, z|C, \lambda) \propto \delta \left( \sum_{j=1}^{p} z_j - pC \right) \delta \left( \sum_{j=1}^{p} w_j^2 - \frac{pC}{\lambda} \right)
\]

is almost equivalent to the mixture prior in high dimensions.

- This proves useful for developing the message passing algorithm.
Factor Graph Representation
Belief Propagation
Factor to Variable Messages

\[ \hat{M}_{n \to \ell}^{t+1} (v_{\ell}, z_{\ell}) \propto \int \prod_{j \neq \ell} dv_j \sum_{z \setminus z_{\ell}} f_n(y_n, z, v) \prod_{j \neq \ell} M_{j \to n}^t(v_j, z_j) \]
Variable to Factor Messages

\[ \mathcal{M}_{\ell \rightarrow n}^t (v_\ell, z_\ell) \propto P(v_\ell, z_\ell) \prod_{m \neq n} \hat{\mathcal{M}}_{m \rightarrow \ell}^t (v_\ell, z_\ell) \]
After $t$ iterations, the approximate posterior marginal belief is:

$$p^t(z_\ell, v_\ell | Y) = \frac{p(z_\ell, v_\ell) \prod_{m=1}^{N} \hat{\mathcal{M}}_{m \rightarrow \ell}^t (v_\ell, z_\ell)}{\int dv_\ell \sum_{z_\ell} p(z_\ell, v_\ell) \prod_{m=1}^{N} \hat{\mathcal{M}}_{m \rightarrow \ell}^t (v_\ell, z_\ell)}$$

where $p(z_\ell, v_\ell)$ is the prior.
Unfortunately,

\[ \mathcal{M}_{n \to \ell}^{t+1} (v_\ell, z_\ell) \propto \int \prod_{j \neq \ell} dv_j \sum_{z \setminus z_\ell} f_n (y_n, z, v) \prod_{j \neq \ell} \mathcal{M}_{j \to n}^t (v_j, z_j) \]

is hard to compute.

Belief propagation is not expected to converge for dense graphical models.
1. Exploit the high-dimensionality:

   Use a **Gaussian approximation**.

2. Impose consistency requirements:

   Use the constraint-based prior to **enforce sparsity and length constraints** self-consistently at each iteration.

Gaussian Approximation (1)

Notice that likelihood factors may be written as:

\[
    f_n(y_n, z, v) = \frac{1}{\sqrt{(2\pi)^p \left(1 + \|w\|^2\right)}} \exp \left(-\frac{y_n^T y_n + \Delta_n^2}{2}\right),
\]

with \(\Delta_n\) defined by:

\[
    \Delta_n = \frac{\sum_{j=1}^{p} y_j^n z_j v_j}{\sqrt{1 + \|w\|^2}}
\]

For large dimension, \(p\), Central Limit Theorem permits a Gaussian approximation.
Gaussian Approximation (2)

For constant $\|\mathbf{w}\|^2$, we replace $\Delta_n$ by:

$$
\frac{y^n_{\ell} z_{\ell} v_{\ell}}{\sqrt{1 + Cp/\lambda}} + \frac{1}{\sqrt{1 + Cp/\lambda}} \sum_{j \neq \ell} y_j^n m_{j \rightarrow n}^t + \sqrt{V_{n \setminus \ell}^t} u
$$

$$
\langle \Delta_{n \setminus \ell} \rangle_{n \setminus \ell}^t
$$

where $u \sim \mathcal{N}(0, 1)$.

$m_{j \rightarrow n}^t$ is the mean of $z_j v_j$ under the cavity distribution with the $n^{th}$ data point removed.
The variance, $V_{n\backslash \ell}^t$ is given by:

$$\frac{1}{1 + Cp/\lambda} \sum_{j,k \neq \ell} y_j^n y_k^n \langle (z_j v_j - m^t_{j \rightarrow n}) (z_k v_k - m^t_{k \rightarrow n}) \rangle^{t}_{n\backslash \ell}$$

For large dimension, $p$, fluctuations about the sample mean are $O\left(\frac{1}{\sqrt{p}}\right)$: $V_{n\backslash \ell}^t$ is *self-averaging*.

$$V^t \approx \frac{1}{(1 + Cp/\lambda)} \left( Cp/\lambda - \sum_{j=1}^{p} (m_j^t)^2 \right)$$
The spike and slab prior can be written:

\[ P(v, z) \propto \prod_{j=1}^{p} \exp \left( -\frac{1}{2} \left( 1 - z_j + G z_j \right) v_j^2 + \gamma z_j \right) \]

where \( \gamma = \ln \left( \frac{C \sqrt{\lambda}}{1 - C} \right) \) and \( G = \lambda \).

Adjust \( G \) and \( \gamma \) at each iteration to satisfy the constraint-based prior on average:

\[ \sum_{j=1}^{p} \langle z_j \rangle^t = Cp \quad \text{and} \quad \sum_{j=1}^{p} \langle z_j v_j^2 \rangle^t = Cp/\lambda \]

Note, after convergence, \( G \neq \lambda \) and \( \gamma \neq \ln \left( \frac{C \sqrt{\lambda}}{1 - C} \right) \).

Consistent with replica analysis.
The spike and slab prior can be written:

\[ P(v, z) \propto \prod_{j=1}^{p} \exp \left( -\frac{1}{2} (1 - z_j + Gz_j) v_j^2 + \gamma z_j \right) \]

where \( \gamma = \ln \left( \frac{C\sqrt{\lambda}}{1-C} \right) \) and \( G = \lambda \).

Adjust \( G \) and \( \gamma \) at each iteration to satisfy the constraint-based prior on average:

\[ \sum_{j=1}^{p} \left\langle z_j \right\rangle^t = Cp \quad \text{and} \quad \sum_{j=1}^{p} \left\langle z_j v_j^2 \right\rangle^t = Cp/\lambda \]

Note, after convergence, \( G \neq \lambda \) and \( \gamma \neq \ln \left( \frac{C\sqrt{\lambda}}{1-C} \right) \).

Consistent with replica analysis.
The spike and slab prior can be written:

\[
P(v, z) \propto \prod_{j=1}^{p} \exp \left( -\frac{1}{2} (1 - z_j + Gz_j) v_j^2 + \gamma z_j \right)
\]

where \( \gamma = \ln \left( \frac{C \sqrt{\lambda}}{1 - C} \right) \) and \( G = \lambda \).

Adjust \( G \) and \( \gamma \) at each iteration to satisfy the constraint-based prior on average:

\[
\sum_{j=1}^{p} \langle z_j \rangle^t = Cp \quad \text{and} \quad \sum_{j=1}^{p} \langle z_j v_j^2 \rangle^t = Cp/\lambda
\]

Note, after convergence, \( G \neq \lambda \) and \( \gamma \neq \ln \left( \frac{C \sqrt{\lambda}}{1 - C} \right) \).

Consistent with replica analysis.
The spike and slab prior can be written:

$$P(v, z) \propto \prod_{j=1}^{p} \exp \left( -\frac{1}{2} \left(1 - z_j + Gz_j \right) v_j^2 + \gamma z_j \right)$$

where $\gamma = \ln \left( \frac{C\sqrt{\lambda}}{1-C} \right)$ and $G = \lambda$.

Adjust $G$ and $\gamma$ at each iteration to satisfy the constraint-based prior on average:

$$\sum_{j=1}^{p} \langle z_j \rangle^t = Cp \text{ and } \sum_{j=1}^{p} \langle z_j v_j^2 \rangle^t = Cp/\lambda$$

Note, after convergence, $G \neq \lambda$ and $\gamma \neq \ln \left( \frac{C\sqrt{\lambda}}{1-C} \right)$.

Consistent with replica analysis.
Replica Analysis (1)

- Compute average of the log marginal likelihood over all possible datasets for $p \to \infty$

- $\alpha = N/p$ is held constant (where $N$ is the sample size).

- Works well for $\alpha \ll 1$ – ‘large $p$ small $N$’

- Not mathematically rigorous, but a useful tool.
Replica Analysis (2)

- Derive expressions involving the posterior mean $w^{PM}$ parameter vector, $w^{PM}$:
  - squared length, $\|w^{PM}\|^2$
  - overlap with the true parameter vector, $w^{PM} \cdot w^t$.
  
  \[
  w^t \sim \prod_{j=1}^{P} \left[ (1 - C_t)\delta(w_j) + C_t \mathcal{N}(w_j|0,(\lambda_t)^{-1}) \right]
  \]

- Can show that the algorithm is consistent with this analysis.

- Can compare algorithm performance to theory using
  \[
  \rho^{PM} = \frac{w^{PM} \cdot w^t}{\|w^{PM}\| \|w^t\|}.
  \]
Simulated Data - DMP vs Theory

DMP

\[ \rho^{PM} \] cosine angle between \( w^{PM} \) and \( w^t \).

\( C \) - fraction of non-zero parameters;

\( N = 200 \) samples, \( \alpha = N/p \);

True Sparsity

Results averaged over 50 sample datasets.

Gibbs
Simulated Data - DMP vs emPCA

**DMP**

- $C$ - fraction of non-zero parameters;
- $N = 200$ samples, $\alpha = N/p$;

**emPCA**

- $\rho_{PM}^C$ cosine angle between $w^{PM}$ and $w^t$.
- Results averaged over 50 sample datasets.
Simulated data - DMP vs SPCA

DMP

\[ \alpha = 0.25 \]

True Sparsity

\[ C \text{ - fraction of non-zero parameters; } \]

\[ N = 200 \text{ samples, } \alpha = N/p; \]

\[ \rho^{PM} \text{ cosine angle between } w^{PM} \text{ and } w^t. \]

Results averaged over 50 sample datasets.
Gene Expression Data - DMP vs emPCA and SPCA

Armstrong *et al.*

Ramaswamy *et al.*

\[ p = 12582, \; N = 72 \]

\[ p = 16063, \; N = 144 \]
Marginal Likelihood Estimation - Simulated Data

\[ C \] - fraction of non-zero parameters;
\[ \lambda \] - assumed signal precision;
True sparsity - 0.1.

\[ N = 200 \text{ samples}; \text{ dimension, } p = 2000. \]
\[ \lambda^t \] - true signal precision.
Summary

- Novel message passing algorithm for Sparse Bayesian PCA in high dimensions
- Message updates rendered tractable using a Gaussian approximation
- Convergence achieved by imposing consistency requirements derived from statistical mechanics analysis.
- Inference of posterior marginals exhibits near optimal performance compared to theory.
- Outperforms two other recently published algorithms.
- Approximation to Marginal Likelihood also available.
Summary

- Novel message passing algorithm for Sparse Bayesian PCA in high dimensions
- Message updates rendered tractable using a Gaussian approximation
- Convergence achieved by imposing consistency requirements derived from statistical mechanics analysis.
- Inference of posterior marginals exhibits near optimal performance compared to theory.
- Outperforms two other recently published algorithms.
- Approximation to Marginal Likelihood also available.
Summary

- Novel message passing algorithm for Sparse Bayesian PCA in high dimensions
- Message updates rendered tractable using a Gaussian approximation
- Convergence achieved by imposing consistency requirements derived from statistical mechanics analysis.
- Inference of posterior marginals exhibits near optimal performance compared to theory.
- Outperforms two other recently published algorithms.
- Approximation to Marginal Likelihood also available.
Summary

- Novel message passing algorithm for Sparse Bayesian PCA in high dimensions
- Message updates rendered tractable using a Gaussian approximation
- Convergence achieved by imposing consistency requirements derived from statistical mechanics analysis.
- Inference of posterior marginals exhibits near optimal performance compared to theory.
- Outperforms two other recently published algorithms.
- Approximation to Marginal Likelihood also available.
Summary

- Novel message passing algorithm for Sparse Bayesian PCA in high dimensions
- Message updates rendered tractable using a Gaussian approximation
- Convergence achieved by imposing consistency requirements derived from statistical mechanics analysis.
- Inference of posterior marginals exhibits near optimal performance compared to theory.
- Outperforms two other recently published algorithms.
- Approximation to Marginal Likelihood also available.
Summary

- Novel message passing algorithm for Sparse Bayesian PCA in high dimensions
- Message updates rendered tractable using a Gaussian approximation
- Convergence achieved by imposing consistency requirements derived from statistical mechanics analysis.
- Inference of posterior marginals exhibits near optimal performance compared to theory.
- Outperforms two other recently published algorithms.
- Approximation to Marginal Likelihood also available.
Hyperparameter estimation using Marginal likelihood.

Extension to multiple factors:
- Relatively straightforward for orthogonal factors.
  (but will require efficient hyperparameter estimation).
- For non-orthogonal factors the best approach is a subject of on-going research.
Hyperparameter estimation using Marginal likelihood.

Extension to multiple factors:
- Relatively straightforward for orthogonal factors.
  (but will require efficient hyperparameter estimation).
- For non-orthogonal factors the best approach is a subject of on-going research.
Explore further

Matlab code available from: http://www.cs.man.ac.uk/~sharpk