Journal of Machine Learning Research 11 (2010) 1771-1798 Submitted 5/09; Revised 1/10; Published 5/10

Bayesian Learning in Sparse Graphical Factor Models via Variational Mean-Field Annealing

Ryo Yoshida YOSHIDAR@ISM.AC.JP
Department of Statistical Modeling
Institute of Statistical Mathematics
Tachikawa, Tokyo 190-8562, Japan

Mike West MW@STAT.DUKE.EDU
Department of Statistical Science
Duke University
Durham, NC 27708-0251, USA

Editor: Michael Jordan

Abstract

We describe a class of sparse latent factor models, called graphical factor models (GFMs), and relevant sparse learning algorithms for posterior mode estimation. Linear, Gaussian GFMs have *sparse, orthogonal* factor loadings matrices, that, in addition to sparsity of the implied covariance matrices, also induce conditional independence structures via zeros in the implied precision matrices. We describe the models and their use for robust estimation of sparse latent factor structure and data/signal reconstruction. We develop computational algorithms for model exploration and posterior mode search, addressing the hard combinatorial optimization involved in the search over a huge space of potential sparse configurations. A mean-field variational technique coupled with annealing is developed to successively generate "artificial" posterior distributions that, at the limiting temperature in the annealing schedule, define required posterior modes in the GFM parameter space. Several detailed empirical studies and comparisons to related approaches are discussed, including analyses of handwritten digit image and cancer gene expression data.

Keywords: annealing, graphical factor models, variational mean-field method, MAP estimation, sparse factor analysis, gene expression profiling

1. Introduction

Bayesian sparse modelling in multivariate analysis is of increasing interest in applications as diverse as life science, economics and information science, and is driving a need for effective computational methods for learning model structure, that is, sparse configurations. Parallel developments of sparse latent factor models (e.g., West, 2003; Griffiths and Ghahramani, 2006; Lucas et al., 2006; Wang et al., 2007; Archambeau and Bach, 2009; Carvalho et al., 2008; Guan and Dy, 2009; Rai and Daumé, 2009) and inherently sparsely structured graphical models (e.g., Jordan, 1999, 2004; Dobra et al., 2004; Jones et al., 2005; Carvalho and West, 2007) have explored Bayesian computations using a range of stochastic and deterministic search methods. With a view to scaling to higher dimensions and identification of regions of interest in model structure space, efficient and effective computation remains a challenge. We describe a previously undeveloped class of sparse graphical factor models (GFMs)—a subclass of linear, Gaussian latent factor models with sparse factor loadings that also induce sparse conditional independencies. In this context, we develop a compu-

tational technique for posterior mode evaluation using a hybrid of variational mean-field method (Attias, 1999; Wainwright and Jordan, 2008) and annealing-based optimization.

As a previously unexplored class of sparse (linear, Gaussian) factor models, the intrinsic graphical structure of the GFM arises from use of an orthogonal factor loadings matrix and appropriate scaling of its columns, together with the usual diagonal covariance matrix for latent factors (with no loss of generality). We show that this generally induces zero elements in the precision matrix of the GFM, as well as the covariance matrix. Particularly, the zero entries in the covariance matrix have corresponding zeros in the precision matrix. We also show that covariance matrices of fitted values (i.e., "data reconstructions") from such a model have the same sparse structure, and demonstrate aspects of robustness of the model in inferring variable-latent factor relationships in the presence of outliers. These properties are not shared in general by sparse factor models that lack the graphical structure on variables, nor of course by non-sparse approaches. These intrinsic properties of the GFM, along with relationships with earlier studies on sparse factor analyses, are discussed in Section 2.

Our *variational mean-field annealing algorithm (VMA2)* addresses the combinatorial optimization involved in aiming to compute approximate posterior modes for GFM parameters in the context of the huge space of zero/non-zero potential patterns in factor loadings. Using a prescribed schedule of decreasing temperatures, VMA2 successively generates tempered "artificial" posteriors that, at the limiting zero temperature, yield posterior modes for both GFM parameters and the 0/1 loadings indicators. Defined via an artificial, dynamic regularization on the posterior entropy of configured sparse structures, VMA2 is developed in Section 3.

Section 4 provides additional algorithmic details, including prior modelling for evaluating degree of sparseness, and a stochastic variant of VMA2 for higher-dimensional problems is described in Section 5. Performance and comparisons on artificial data appear in Section 6. Section 7 summarizes extensive, detailed empirical comparisons with related approaches in analyses of hand-written digit images and cancer gene expression data. Section 8 concludes with brief additional comments. A range of detailed supplementary materials, extended discussion on the gene expression studies and `R` code, is accessible from `http://daweb.ism.ac.jp/~yoshidar/anneals/`.

2. Sparse Graphical Factor Models

We describe the GFM with some intrinsic graphical properties, followed by connections to previously developed classes of sparse latent factor analyses.

2.1 GFM Form

Observed sample vectors $x_i \in \mathbb{R}^p$ in p dimensional feature space are each linearly related to independent, unobserved Gaussian latent factor vectors $\lambda_i \in \mathbb{R}^k$ with additional Gaussian noise. We are interested in sparse variable-factor relationships so that the bipartite mapping $\lambda \to x$ is sparse, with the underlying $p \times k$ matrix of coefficients—the *factor loadings matrix*—having a number of zero elements; the $p \times k$ binary matrix Z defines this *configured sparsity pattern*. We use a sparse, orthogonal loading matrix and diagonal covariance matrices for both latent factors and residuals; the model is mathematically identified in the usual sense in factor analysis (Anderson, 2003).

With Z as the $p \times k$ binary matrix with elements z_{gj} such that variable g is related to factor j if and only if $z_{gj} = 1$, the GFM is

$$x_i = \Psi^{1/2}\Phi_Z\lambda_i + \nu_i \quad \text{with } \lambda_i \sim \mathcal{N}(\lambda_i|0,\Delta) \text{ and } \nu_i \sim \mathcal{N}(\nu_i|0,\Psi)$$

where: *(a)* the factor loading matrix $\Psi^{1/2}\Phi_Z$ has $\Phi_Z \equiv \Phi \circ Z$ with $\circ$ representing element-wise product; *(b)* Φ_Z is orthogonal, that is, $\Phi_Z'\Phi_Z = I_k$; *(c)* the factors have diagonal covariance matrix $\Delta = \text{diag}(\delta_1,\ldots,\delta_k)$; and *(d)* the idiosyncratic Gaussian noise (or residual) ν_i is independent of λ_i and has covariance matrix $\Psi = \text{diag}(\psi_1,\ldots,\psi_p)$. The implied covariance matrix of the sampling model, Σ, and the corresponding *precision matrix*, Σ^{-1}, are

$$\Sigma = \Psi^{1/2}\{I + \Phi_Z\Delta\Phi_Z'\}\Psi^{1/2} \quad \text{and} \quad \Sigma^{-1} = \Psi^{-1/2}\{I - \Phi_Z T\Phi_Z'\}\Psi^{-1/2} \tag{1}$$

where $T = \text{diag}(\tau_1,\ldots,\tau_k)$ with $\tau_j = \delta_j/(1+\delta_j)$ $(j = 1:k)$. In general, sparse loading matrices induce some zero elements in the covariance matrix whether or not they are orthogonal, but *not* in the implied precision matrix. In the GFM here, however, a sparse factor model also induces off-diagonal zeros in Σ^{-1}. Zeros in the precision matrix defines a conditional independence or graphical model, hence the GFM terminology. In (1), the pattern of sparsity (location of zero entries) in the covariance and precision matrices are the same. The set of variables associated with one specific factor forms a clique in the induced graphical model, with sets of variables that have non-zero loadings on any two factors lying in the separating subgraph between the corresponding cliques. Hence, we have a natural and appealing framework in which sparse factor models and graphical models are reconciled and consistent.

2.2 Some Model Attributes

In general, a non-orthogonal factor model with the sparse loading matrix W—a sparse extension of probabilistic PCA (Bishop, 1999, 2006)—has the form

$$x_i = W\lambda_i + \nu_i \text{ with } \lambda_i \sim N(0,I) \text{ and } \nu_i \sim N(0,\Psi).$$

The GFM arises when a singular value decomposition is applied to the scaled-factor loading matrix $\Psi^{-1/2}W = \Phi_Z\Delta^{1/2}R$ with a $k \times k$ orthogonal matrix R being removed. This non-orthogonal model defines a Bayes optimal reconstruction of the data via the fitted values (or extracted signal)

$$\hat{x}(x_i) := W\mathbb{E}[\lambda_i|x_i] = WW'(WW'+\Psi)^{-1}x_i.$$

Then, asymptotically,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{x}(x_i)\hat{x}(x_i)' \xrightarrow{p} \mathbb{C}ov[\hat{x}(x_i)] = WW'(WW'+\Psi)^{-1}WW'$$

and this is generally a non-sparse matrix (no zero entries) even though W is sparse. This is an inconsistency in the sense that data reconstructions should be expected to share the dominant patterns of covariance sparsity evident in the original covariance matrix $\mathbb{C}ov[x_i] = WW'+\Psi$. In the GFM, however, $\mathbb{C}ov[\hat{x}(x_i)] = \Psi^{1/2}\Phi_Z G\Phi_Z'\Psi^{1/2}$ where G is diagonal with entries $\delta_j^2/(1+\delta_j)$. In such cases, $\mathbb{C}ov[\hat{x}(x_i)]$ is sparse and shares the same 0 elements as $\mathbb{C}ov[x_i]$.

Another feature of the GFM is related to a robust property acquired by the implied graphical structure. Consider an example of 4 variables $x_i' = (x_{i1},x_{i2},x_{i3},x_{i4})$ and 2 factors $\lambda_i' = (\lambda_{i1},\lambda_{i2})$

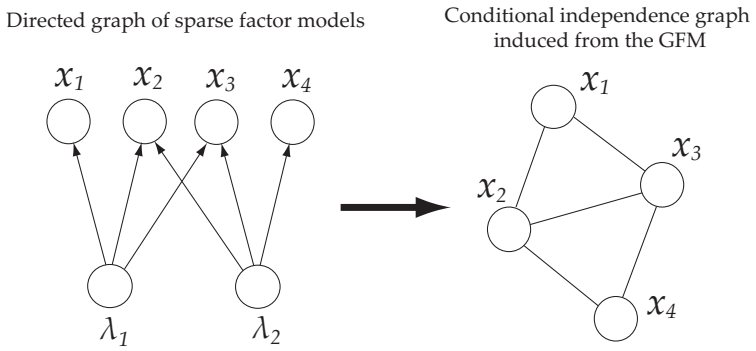

Figure 1: Graphical model structure of an example GFM.

with two cliques in the conditional independence graph; $\{x_{i1}, x_{i2}, x_{i3}\} \leftarrow \lambda_{i1}$ and $\{x_{i2}, x_{i3}, x_{i4}\} \leftarrow \lambda_{i2}$ (see Figure 1). The graph defines the decomposition of the joint density $p(x_{i1}, x_{i2}, x_{i3}, x_{i4}) = p(x_{i1}|x_{i2}, x_{i3})p(x_{i2}, x_{i3}|x_{i4})p(x_{i4})$ or $p(x_{i1}, x_{i2}, x_{i3}, x_{i4}) = p(x_{i4}|x_{i2}, x_{i3})p(x_{i2}, x_{i3}|x_{i1})p(x_{i1})$. This implies that presence of one or more outliers in the isolated feature variable, that is, x_{i1} or x_{i4}, associated with a single factor clique, has no effect on the variables, x_{i4} or x_{i1}, once the intermediate variables x_{i2} and x_{i3} are given. Then, the parameters involved in $p(x_{i1})$ or $p(x_{i4})$, for instance, the loading components and the noise variances corresponding to the isolated variable, can be estimated independently of the impact of outliers in x_{i4} or x_{i1}. The numerical experiment shown in Section 7.1 highlights this robustness property in terms of data compression/restoration tasks, with comparison to other sparse factor models.

2.3 Likelihood, Priors and Posterior

Denote by Θ the full set of parameters $\Theta = \{\Phi, \Delta, \Psi\}$. Our computations aim to explore model structures Z and corresponding posterior modes of parameters Θ under the posterior $p(Z, \Theta|X)$ using specified priors and based on the n observations forming the columns of the $p \times n$ data matrix X.

2.3.1 Likelihood Function

The likelihood function is

$$p(X|Z, \Theta) \propto |\Psi|^{-n/2}|I - T|^{n/2}\text{etr}(-S\Psi^{-1}/2 + \Psi^{-1/2}S\Psi^{-1/2}\Phi_Z T\Phi_Z'/2) \tag{2}$$

where $\text{etr}(A) = \exp(\text{trace}(A))$ for any square matrix A, and S is the sample sum-of-square matrix $S = XX'$ with elements s_{gh}. In (2), the factor loadings appear only in the last term and form the important statistic

$$\text{trace}(\Psi^{-1/2}S\Psi^{-1/2}\Phi_Z T\Phi_Z') = \sum_{j=1}^{k} \tau_j \phi_{z_j}' \Psi^{-1/2}S\Psi^{-1/2}\phi_{z_j}$$

where ϕ_{z_j} is column j of Φ_Z, or $\phi_{z_j} = \phi_j \circ z_j$ where ϕ_j is column j of Φ and z_j is column j of Z.

2.3.2 PRIORS ON Θ AND Z

Priors over non-zero factor loadings may reflect substantive *a priori* knowledge if available, and will then be inherently context specific. For examples here, however, we use uniform priors $p(\Theta|Z)$ for exposition. Note that, on the critical factor loadings elements Φ, this involves a uniform on the hypersphere defined by the orthogonality constraint that is then simply conditioned (by setting implied elements of Φ to zero) as we move across candidate models Z.

Concerning the sparse structure Z, we adopt independent priors on the binary variates z_{gj} with $\text{logit}(\Pr(z_{gj}=1|\zeta_{gj})) = -\zeta_{gj}/2$ where $\text{logit}(p) = \log(p/(1-p))$ and the parameters ζ_{gj} are assigned hyperpriors and included in the overall parameter set in later. Beta priors are obvious alternatives to this; the logit leads to a minor algorithmic simplification, but otherwise the choice is arbitrary. Using beta priors can be expected to lead to modest differences, if any of practical relevance, in many cases, and users are free to explore variants. The critical point is that including Bayesian inference on these $p \times k$ sparsity-determining quantities leads to "self-organization" as their posterior distributions concentrate on larger or smaller values. Examples in Section 6 highlight this.

2.4 MAP Estimation for (Θ, Z) in GFMs

Conditional on the $p \times k$ matrix of sparsity control hyperparameters ζ whose elements are the ζ_{gj}, it follows that posterior modes (Z, Θ) maximize

$$\begin{aligned} 2\log p(Z,\Theta|X,\zeta) &= 2\log p(\Theta|Z) - \sum_{g=1}^{p}\sum_{j=1}^{k} z_{gj}\zeta_{gj} - \sum_{g=1}^{p}(n\log\psi_g + s_{gg}\psi_g^{-1}) \\ &\quad + \sum_{j=1}^{k}(n\log(1-\tau_j) + \tau_j\phi_{zj}'\Psi^{-1/2}S\Psi^{-1/2}\phi_{zj}). \end{aligned} \tag{3}$$

The first two terms in (3) arise from the specified priors for Θ and Z, respectively. The quadratic form in the last term is $\phi_{zj}'\Psi^{-1/2}S\Psi^{-1/2}\phi_{zj} = \phi_j'S(z_j,\Psi)\phi_j$ for each j, where the key $p \times p$ matrices $S(z_j,\Psi)$ have elements $(S(z_j,\Psi))_{gh}$ given by

$$(S(z_j,\Psi))_{gh} = z_{gj}z_{hj}s_{gh}(\psi_g\psi_h)^{-1/2}, \qquad \text{for } g,h = 1:p. \tag{4}$$

The (relative) signal-to-noise ratios $\tau_j = \delta_j/(1+\delta_j)$ control the roles played by the last term in (3).

Optimizing (3) over Θ and Z involves many discrete variables and the consequent combinatorial computational challenge. Greedy hill-climbing approaches will get stuck at improper local solutions, often and quickly. The VMA2 method in Section 3 addresses this.

2.5 Links to Previous Sparse Factor Modelling and Learning

In the MAP estimation defined by (3), there are evident connections with traditional sparse principal component analyses (sparse PCA; Jolliffe et al., 2003, Zou et al., 2006 and d'Aspremont et al., 2007). If $\Psi = I$ and $\Delta = I$, the latter likelihood component in (3) is the pooled-variance of projections, that is, $\sum_{j=1}^{k}\phi_j'S(z_j,I)\phi_j$, constructed by the k sparse loading vectors. This is the central statistic optimized in many sparse PCAs. Differences among existing sparse PCAs arise in the way they regulate degrees of sparseness and whether or not orthogonality is imposed on the loading vectors.

The direct sparse PCA of d'Aspremont et al. (2007) imposes an upper-bound $d > 0$ on the cardinality of z_j (the number of non-zero elements), with a resulting semidefinite programming of computational complexity $O(p^4\sqrt{log(p)})$. The applicability of that approach is therefore limited to problems with p rather small. Such cardinality constraints can be regarded as suggestive of structure for the prior distribution on ζ in our model.

The SCoTLASS algorithm of Jolliffe et al. (2003) uses ℓ_1-regularization on loading vectors, later extended to SPCA using elastic nets by Zou et al. (2006). Recently, Mairal et al. (2009) presented a ℓ_1-based dictionary learning for sparse coding in which the method aims to explore sparsity on factor-sample mapping rather than that on factor-variable relations. Setting Laplace-like prior distributions on scale loadings is a counterpart of ℓ_1-based penalization (Jolliffe et al., 2003; Zou et al., 2006). However, our model-based perspective aims for a more probabilistic analysis, with advantages in probabilistic assessment of appropriate dimension of the latent factor space as well as flexibility in the determination of effective degrees of sparseness via the additional parameters ζ. Other than the preceding studies, ℓ_1-regularizations have widely been employed to make sparse latent factor analyses. Archambeau and Bach (2009) developed a general class of sparse latent factor analyses involving sparse probabilistic PCA (Guan and Dy, 2009) and a sparse variant of probabilistic canonical correlation analysis. A key idea of Archambeau and Bach (2009) is to place the automatic relevance determination (ARD) prior of Mackay (1995) on each loading component, and to apply a variational mean-field learning method.

Key advances in Bayesian sparse factor analysis build on non-parametric Bayesian modelling in Griffiths and Ghahramani (2006) and Rai and Daumé (2009), and developments in Carvalho et al. (2008) stemming from the original sparse Bayesian models in West (2003). Carvalho et al develop MCMC and stochastic search methods for posterior exploration. MCMC posterior sampling can be effective but is hugely challenged as the dimensions of data and factor variables increase. Our focus here is MAP evaluation with a view to scaling to increasingly large dimensions, and we leave open the opportunities for future work on MCMC methods in GFMs.

Most importantly, as remarked in Section 2.2, the GFM differs from some of the forgoing models in the conditional independence graphical structures induced. This characteristic contributes to preserving sparse structure in the data compression/reconstruction process and also to the outlier robustness issue. We leave further comparative discussion to Section 7.1, where we evaluate some of the foregoing methods relative to the sparse GFM analysis in an image processing study.

3. Variational Mean-Field Annealing for MAP Search

Finding MAP estimates of the augmented posterior distribution (3) involves many discrete variables z_{gj}. Then, commonly applied search methods such as greedy hill-climbing algorithm often get stuck in improper local solutions. Here, we present a general framework of VMA2 enabling us to escape local mode traps by exploiting annealing.

3.1 Basic Principle

Relative to (3), consider the class of extended objective functions

$$\mathcal{G}_T(\Theta,\omega) = \sum_{Z\in\mathscr{Z}} \omega(Z)\log p(X,Z,\Theta|\zeta) - T\sum_{Z\in\mathscr{Z}} \omega(Z)\log\omega(Z) \tag{5}$$

where $\omega(Z)$—the *sparsity configuration probability*—represents *any distribution* over $Z \in \mathscr{Z}$ that may depend on (X, Θ, ζ), and where $T \geq 0$. This modifies the original criterion (3) by taking the expectation of $p(X, Z, \Theta|\zeta)$ with respect to $\omega(Z)$—the expected complete data log-likelihood in the context of EM algorithm—and by the inclusion of Shannon's entropy of $\omega(Z)$ with the *temperature* multiplier T.

Now, view (5) as a criterion to maximize over (Θ, ω) jointly for any given T. The following is a key result:

Proposition 1 *For any given parameters* Θ *and temperature* T*, (5) is maximized with respect to* ω *at*

$$\omega_T(Z) \propto p(Z|X, \Theta, \zeta)^{1/T}. \tag{6}$$

Proof See the Appendix. ■

For any given Θ, a large T leads to $\omega_T(Z)$ being rather diffuse over sparse configurations Z so that iterative optimization—alternating between Θ and ω—will tend to move more easily and freely around the high-dimensional space Z. This suggests annealing beginning with the temperature T large and successively reducing towards zero. We note that:

- As $T \to 0$, $\omega_T(Z)$ converges to a distribution degenerate at the conditional mode $\hat{Z}(\Theta, \zeta)$ of $p(Z|X, \Theta, \zeta)$, so that
- joint maximization of $\mathcal{G}_T(\Theta, \omega)$ would approach the *global maximum of the exact posterior* $p(\Theta, Z|X, \zeta)$ as $T \to 0$.

The notion of the annealing operation is to realize a gradual move of successively-generated solutions for Θ and $\omega_T(Z)$, and to escape local mode traps by exploiting annealing. Note that, for any given tempered posterior (6), the expectation in the first term of (5) is virtually impossible to be taken due to the combinatorial explosion. In what follows, we introduce VMA2 as a mean-field technique coupled with the annealing-based optimization to overcome this central computational difficulty.

3.2 VMA2 based on Factorized, Tempered Posteriors

To define and implement a specific algorithm, we constrain the otherwise arbitrary *"artificial configuration probabilities"* ω, and do so using a construction that induces analytic tractability. We specify the simplest, factorized form

$$\omega(Z) = \prod_{g=1}^{p} \prod_{j=1}^{k} \omega(z_{gj}) := \prod_{g=1}^{p} \prod_{j=1}^{k} \omega_{gj}^{z_{gj}} (1 - \omega_{gj})^{1 - z_{gj}}$$

in the same way as conventional Variational Bayes (VB) procedures do. In this GFM context, the resulting optimization is eased using this independence relaxation as it gives rise to tractability in computing the conditional expectation in the first term of (5).

If $T = 1$, and given the factorized ω, the objective function $\mathcal{G}_1$ exactly agrees with the *free energy*, which bounds the posterior marginal as

$$\log \sum_{Z \in \mathscr{Z}} p(X, \Theta, Z|\zeta) \geq \mathcal{G}_1(\Theta, \omega).$$

The lower-bound $\mathcal{G}_1$ is the criterion that the conventional VB methods aim to maximize (Wainwright and Jordan, 2008). This indicates that any solutions corresponding to the VB inference can be obtained by stopping the cooling schedule at $T = 1$ in our method. Similar ideas have, of course, been applied in deterministic annealing EM and annealed VB algorithms (e.g., Ueda and Nakano, 1998). These methods exploit annealing schemes to escape from local traps during coordinate-basis updates in aiming to define variational approximations of posteriors.

Even with this relaxation, maximization over $\omega(Z)$ cannot be done for all elements of Z simultaneously and so is approached sequentially—sequencing through each ω_{gj} in turn while conditioning the others. For any given T this yields the optimizing value given by

$$\omega_{gj}(T) \propto \exp\Big\{\frac{1}{T}\sum_{\mathscr{Z}_{\mathcal{C}\setminus\{g,j\}}}\prod_{h\neq g}\prod_{l\neq j}\omega(z_{hl})\log p(z_{gj}=1|X,Z_{\mathcal{C}\setminus\{g,j\}},\Theta,\zeta)\Big\} \tag{7}$$

where $\mathcal{C}$ denotes the collection of all indices (g, j) for the p features and k factor variables, $\mathcal{C}\setminus\{g, j\}$ is the set of the paired indices (h,l) such that $(h,l) \neq (g, j)$, and $\mathscr{Z}_{\mathcal{C}\setminus\{g,j\}}$ stands for the set of z_{hl}s other than z_{gj}.

Starting with $\omega_{gj} \simeq 1/2$ at an initial large value of T, (7) gradually concentrates to the point mass as T decays to zero slowly:

$$\hat{z}_{gj} := \lim_{T\downarrow 0}\omega_{gj}(T) = \begin{cases} 1, & \text{if } \displaystyle\sum_{\mathscr{Z}_{\mathcal{C}\setminus\{g,j\}}}\prod_{h\neq g}\prod_{l\neq j}\omega(z_{hl})\log\frac{p(z_{gj}=1,X,Z_{\mathcal{C}\setminus\{g,j\}},\Theta,\zeta)}{p(z_{gj}=0,X,Z_{\mathcal{C}\setminus\{g,j\}},\Theta,\zeta)} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

It remains true that, at the limiting zero temperature, the global maximum of $\mathcal{G}_T(\Theta,\omega)$ is the set of $p \times k$ point masses at the global posterior mode of $p(\Theta, Z|X,\zeta)$. This is seen trivially as follows: *(i)* As $T \to 0$, and with the non-factorized ω in (5), we have limiting value

$$\sup_Z \log p(X,\Theta,Z|\zeta) = \sup_\omega \mathcal{G}_0(\Theta,\omega) \tag{8}$$

with the point mass $\omega(Z) = \delta_{\hat{Z}}(Z)$ at the location of the global maximum $(\hat{Z})_{gj} = \hat{z}_{gj}$. Further, *(ii)* any point mass $\delta_{\hat{Z}}(Z)$ is representable by a fully factorized $p \times k$ point masses as $\delta_{\hat{Z}}(Z) = \prod_{g,j}\delta_{\hat{z}_{gj}}(z_{gj})$.

It is stressed that the coordinate-basis updates (7) cannot, of course, guarantee convergence to the *global* optimum even with prescribed annealing. Nevertheless, VMA2 represents a substantial advance in its ability to move more freely and escape local mode traps. We also note the generality of the idea, beyond factor models and also potentially using penalty functions other than entropy.

4. Sparse Learning in Graphical Factor Models

We first provide a specific form of VMA2 for the GFM, and then address the issue of evaluating relevant degrees of sparseness.

4.1 MAP Algorithm

Computations alternate between conditional maximization steps for ω and Θ while reducing the temperature T. At each step, the value of the objective function (5) is kept to refine until convergence where the temperature reaches to zero. Specifically:

1: Set a cooling schedule $\mathcal{T} = \{T_1, \ldots, T_d\}$ of length d where $T_d = 0$;
2: Set ζ;
3: Initialize Θ;
4: Initialize $\omega(Z)$;
5: $i \leftarrow 0$;
6: `while` $(\{\text{the loop is not converged }\} \wedge \{i \leq d\})$
7: $i \leftarrow i + 1$;
8: Compute configuration probabilities $\omega_{gj}(T_i)$;
9: Optimize with respect to each column ϕ_j $(j = 1 : k)$ of Φ in turn under full-conditioning;
10: Optimize with respect to Δ under full-conditioning;
11: Optimize with respect to Ψ under full-conditioning;
12: Optimize with respect to ζ under full-conditioning;
13: `end while`

We now summarize key components in the iterative, annealed computation defined above.

4.2 Sparse Configuration Probabilities

First consider maximization with respect to each sparse configuration probability ω_{gj} conditional on all others. We note that the first term in (5) involves the expectation over Z with respect to the probabilities ω, denoted by $\mathbb{E}_\omega[\cdot]$. Accordingly, for the key terms $S(z_j, \Psi)$ we have

$$\mathbb{E}_\omega[S(z_j, \Psi)] = \Omega_j \circ (\Psi^{-1/2} S \Psi^{-1/2}) \text{ with } (\Omega_j)_{gh} = \begin{cases} \omega_{gj}, & \text{if } g = h, \\ \omega_{gj}\omega_{hj}, & \text{otherwise.} \end{cases} \tag{9}$$

Introduce the notation $\Psi^{-1/2} S \Psi^{-1/2} = (s_1(\Psi), \ldots, s_p(\Psi))$ to represent the p columns of the scaled-sample sum-of-square matrix here, and define the $p-$vector

$$\tilde{\omega}_{gj} = (\omega_{1j}, \ldots, \omega_{g-1,j}, 1, \omega_{g+1,j}, \ldots, \omega_{pj})'.$$

Then, the partial derivative of (5) with respect to ω_{gj} conditional on Θ and the other configuration probabilities leads to

$$\text{logit}(\omega_{gj}(T)) = H_{gj}(\zeta_{gj})/T \quad \text{where} \quad H_{gj}(\zeta_{gj}) := \tau_j \phi_{gj} (\phi_j \circ \tilde{\omega}_{gj})' s_g(\Psi) - \zeta_{gj}.$$

This directly yields the conditional maximizer for ω_{gj} in terms of the tempered negative energy $H_{gj}(\zeta_{gj})/T$. As the temperature T is reduced towards zero, the resulting estimate tends towards 0 or 1 according to the sign of $H_{gj}(\zeta_{gj})$.

4.3 Conditional Optimization over Φ

The terms in (5) that involve Φ are simply the expectation of the quadratic forms in the last term of (3), with the term for each column ϕ_j involving the key matrices $S(z_j, \Psi)$ defined in (4), for each $j = 1 : k$. At each step through the overall optimization algorithm in Section 4.1, we sequence through these columns of the loadings matrix in turn conditioning on the previously optimized

values of all other columns. In the context of the overall iterative MAP algorithm, this yields global optimization over Φ as $T \to 0$.

Conditional optimization then reduces to the following: for each $j = 1:k$, sequence through each column ϕ_j in turn and at each step

$$\begin{aligned} &\underset{\phi_j}{\text{maximize}} && \phi_j' \mathbb{E}_\omega[S(z_j, \Psi)]\phi_j \\ &\text{subject to} && \phi_j'\phi_j = 1 \quad \text{and} \quad \phi_m'\phi_j = 0 \text{ for } m \neq j, m = 1:k. \end{aligned} \tag{10}$$

The optimization conditions on the most recently updated values of all other columns $m \neq j$ at each step, and is performed as one sweep as the line 9 in the algorithm of Section 4.1. Column order can be chosen randomly or systematically each time while still maintaining convergence. In this step, we stress that the original orthogonality condition is modified to $\Phi_Z'\Phi_Z = I \to \Phi^T\Phi = I$ in (10). It remains the case that iteratively refined estimates obtained from (10) satisfy the original condition at the limiting zero temperature, yielding sparsity for $\mathbb{E}_\omega[S(z_j, \Psi)]$, as detailed in the mathematical derivations in supplementary material.

The specific computations required for the conditional optimization in (10) are as follows (with supporting details in the Appendix). Note that the central matrices $\mathbb{E}_\omega[S(z_j, \Psi)]$ required here are trivially available from Equation (9).

1: Compute the $p \times (k-1)$ matrix $\Phi_{(-j)} = \{\phi_m\}_{m \neq j}$ by simply deleting column j from Φ;

2: Compute the $p \times p$ projection matrix $N_j = I_p - \Phi_{(-j)}\Phi_{(-j)}'$;

3: Compute the eigenvector φ_j corresponding to the most dominant eigenvalue of $N_j \mathbb{E}_\omega[S(z_j, \Psi)]N_j$;

4: Compute the required optimal vector $\phi_j = N_j\varphi_j / \|N_j\varphi_j\|$.

This procedure solves (10) by optimizing over an eigenvector already constrained by the orthogonality conditions. Here N_j spans the null space of the current $k-1$ columns of $\Phi_{(-j)}$, so $N_j \mathbb{E}_\omega[S(z_j, \Psi)]N_j$ defines the projection of $\mathbb{E}_\omega[S(z_j, \Psi)]$ onto the orthogonal space and eigenvectors φ_j lie in the null space. It remains to ensure that the computed value ϕ_j is of unit length, which involves the normalization in the final step in part 4. Selecting the eigenvector with maximum eigenvalue ensures the conditional maximization in (10).

4.4 Conditional Optimization over Δ

The variances δ_j of the latent factors appear in Equations (3) and (5) in the sum over $j = 1:k$ of terms

$$-n\log(1+\delta_j) + \delta_j(1+\delta_j)^{-1}\phi_j' \mathbb{E}_\omega[S(z_j, \Psi)]\phi_j.$$

This is unimodal in δ_j with maximizing value

$$\hat{\delta}_j = \max\{0, n^{-1}\phi_j' \mathbb{E}_\omega[S(z_j, \Psi)]\phi_j - 1\}, \tag{11}$$

and so the update at the line 10 of the MAP algorithm of Section 4.1 computes these values in parallel for each factor $j = 1:k$. Note that this may generate zero values, indicating the removal of the corresponding factors from the model, and so inducing an intrinsic ability to prune the number

of factors as being redundant in a model specified initially with a larger, encompassing value of k. The configured sparse structure drives this pruning; any specific factor j that is inherently very sparse generates a smaller value of the projected "variance explained" $\phi_j' \mathbb{E}_\omega[S(z_j, \Psi)]\phi_j$, and so can lead to $\hat{\delta}_j = 0$ as a result.

4.5 Conditional Optimization over Ψ

The diagonal noise covariance matrix Ψ appears in the objective function of Equation (5) in terms that can be re-expressed as

$$-n \log|\Psi| - \text{trace}(S\Psi^{-1}) + \sum_{j=1}^{k} \tau_j \text{trace}(\phi_j \phi_j' \Psi^{-1/2}(\Omega_j \circ S)\Psi^{-1/2})$$

where $\tau_j = \delta_j/(1+\delta_j)$ for each j. Differentiating this with respect to $\Psi^{-1/2}$ yields the gradient equation:

$$n\text{diag}^{-1}(\Psi^{1/2}) - \text{diag}^{-1}(S\Psi^{-1/2}) + \sum_{j=1}^{k} \tau_j \text{diag}^{-1}(\phi_j \phi_j' \Psi^{-1/2}(\Omega_j \circ S)) = 0,$$

where $\text{diag}^{-1}(A)$ denotes the vector of the diagonal elements in A. Iterative solution of this non-linear equation in Ψ can be performed via the reduced implicit equation

$$\text{diag}^{-1}(\Psi) = n^{-1}\text{diag}^{-1}(\{I_p - \sum_{j=1}^{k} \tau_j (\phi_j \phi_j') \circ (\Psi^{-1/2}\Omega_j \Psi^{1/2})\}S).$$

4.6 Degrees of Sparseness

The prior over the logistic hyperparameters $\zeta = \{\zeta_{gj}\}$ defining the Bernoulli probabilities for the z_{gj} is important in encouraging relevant degrees of sparseness. Extending the model via an hierarchical prior for these parameters enables adaptation to data in evaluating relevant degrees of sparseness. One first class of priors is used here, taking the ζ_{gj} to be conditionally independent and drawn from the prior with positive part Gaussian distribution $N_+(\zeta_{gj}|\mu,\sigma)$ for some specified mean and variance (μ,σ). The annealing search can now be extended to include ζ, simply embedding conditional optimization of (5) under this prior within each step of the iterative search. The conditional independence structure of the model easily yields unique solutions for each of the ζ_{gj} in parallel as values satisfying

$$\omega_{gj} = \frac{\exp(-\zeta_{gj}/2)}{1+\exp(-\zeta_{gj}/2)} - \frac{\zeta_{gj}-\mu}{2\sigma}. \tag{12}$$

Solutions to (12) are trivially, iteratively computed. Evidently, as ω_{gj} approaches 0 or 1, the solution for ζ_{gj} is shifted to the corresponding boundary. ζ_{gj} as a function of ω_{gj} for several values of (μ,σ).

As mentioned earlier, the choice of this logit/truncated normal prior is a subjective preference and could be replaced by others, such as beta priors. Again, we expect that this would typically lead to modest differences, if any of practical relevance, in many cases.

5. A Stochastic Search Variant for Large p

In problems with larger numbers of variables, the computations quickly become challenging, especially in view of the repeated eigen-decompositions required for updating factor loading matrix. In our examples and experiments, analysis with dimensions $p \sim 500$ would be feasible using our own `R` code (`vma2gfm()` available from the supplementary web site), but computation time then rapidly increases with increasing p. More efficient low level coding will speed this, but nevertheless it is of interest to explore additional opportunities for increasing the efficiency of the MAP search.

To reduce the computational time, we explore a stochastic variant of the original deterministic VMA2 that uses realized Z matrices from current, conditional configuration probabilities $\omega_{gj}(T)$ at each stage of the search process. The realized binary matrix $Z = [z_1, \ldots, z_k]$ replaces the full matrix $\mathbb{E}_\omega[S(z_j, \Psi)]$ with a sparse alternative $S(z_j, \Psi)$. In larger, very sparse problems, this will enable us to greatly reduce the computing time as each eigen-decomposition can be computed based only on the components related to non-zero z_{gj} values. This leads to a stochastic annealing search with all other steps unchanged. We also have the additional benefit of the introduced randomness aiding in potentially moving away from the stuck in suboptimal solutions. It should be stressed that this is an optional complement to the deterministic algorithm and one that may be used for an initial period of time prior to enable swifter initial iterations from arbitrary initial values, prior to switching to the deterministic annealing once in the region of a posterior mode.

The modified search procedure over ϕ_j in Equation (10) is:

1. Draw a set of binary values $\hat{z}_{gj}$, $g = 1, \ldots, p$, according to the current configuration probabilities $\omega_{gj}(T)$;
2. Define the set of *active variables* by $\mathcal{A}_j = \{g | g \in 1:p, \hat{z}_{gj} = 1\}$; denote by $\phi_{j,\{\mathcal{A}_j\}}$ the sub-vector of ϕ_j for only the active variables, and $S_{\{\mathcal{A}_j\}}(z_j, \Psi)$ the submatrix of $S(z_j, \Psi)$ whose rows and columns correspond to only the active variables;
3. Solve the reduced optimization conditional on the $\mathcal{A}_j$, via:

$$\begin{aligned} \underset{\phi_{j,\{\mathcal{A}_j\}}}{\text{maximize}} \quad & \phi'_{j,\{\mathcal{A}_j\}} S_{\{\mathcal{A}_j\}}(\hat{z}_j, \Psi) \phi_{j,\{\mathcal{A}_j\}} \\ \text{subject to} \quad & \|\phi_{j,\{\mathcal{A}_j\}}\|^2 = 1 \ \text{ and } \ \phi'_{m,\{\mathcal{A}_j\}} \phi_{j,\{\mathcal{A}_j\}} = 0 \text{ for } m \neq j. \end{aligned}$$

4. Update the full $p-$vector ϕ_j with elements $\phi_{j,\{\mathcal{A}_j\}}$ for the active variables and all other elements zero.

For example, in a problem with $p = 5000$ but sparseness of the order of 5%, the $\mathcal{A}_j$ will involve a few hundred active variables, and eigenvalue decomposition will then be performed on matrices of that order rather than 5000×5000. We note also that this strategy requires a modification to the update operation for the configuration probabilities: the ω_{gj} will be updated at any one step only for the current indices $g \in \mathcal{A}_j$, keeping the remaining z_{gj} at values previously obtained.

6. Experimental Results on Synthetic Data

Performance and comparisons on artificial data are shown to highlight some learning properties of the GFM.

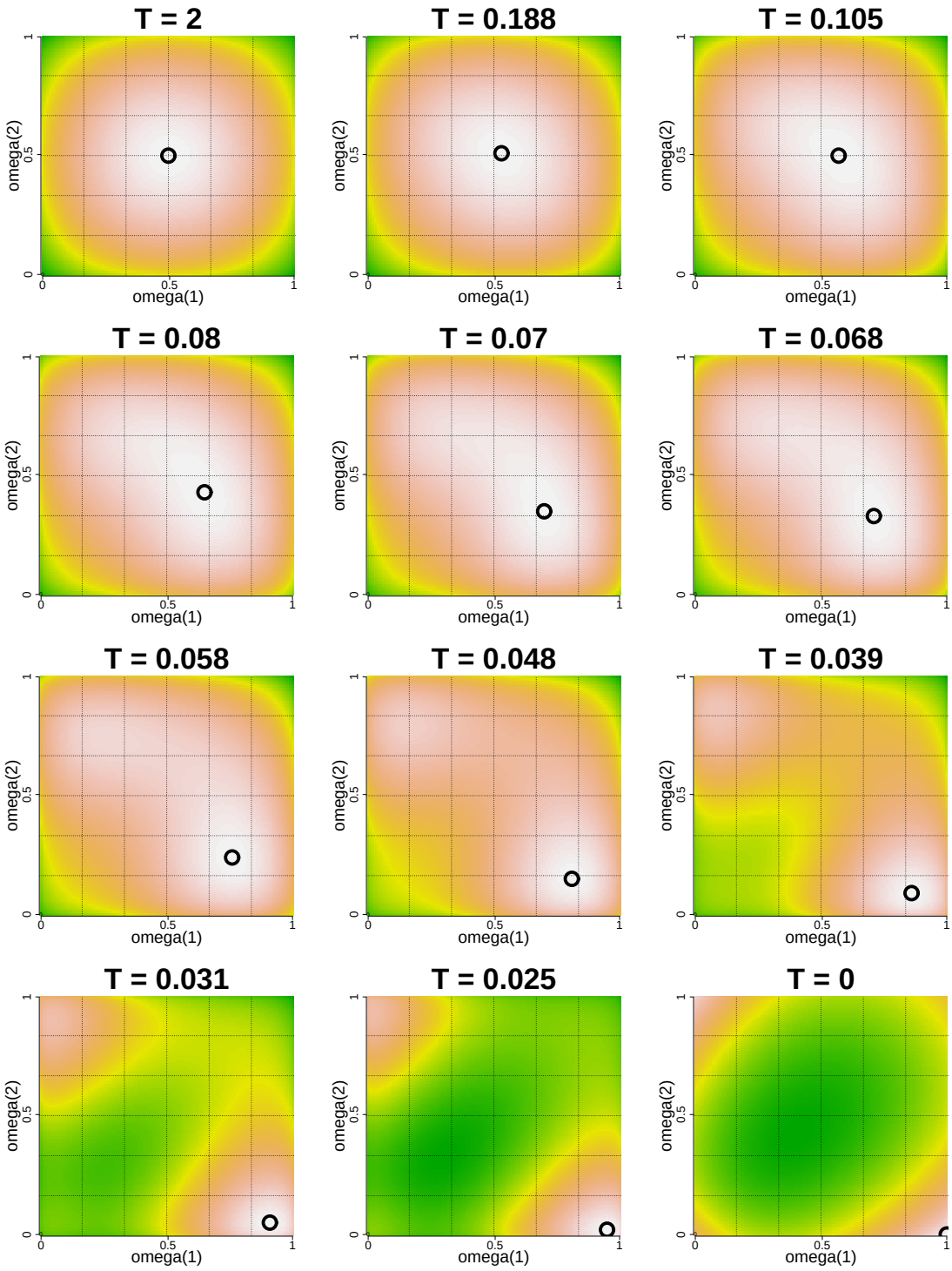

Figure 2: Display of evolving $\mathcal{G}_T(\hat{\Theta}, \omega)$ in the annealing process (from $T = 2$ to $T = 0$) with contour plots. The black circle in each panel indicates the maximum point, and that corresponding to $T = 0$ in the panel on the bottom-right corner indicates the optimal sparse structure.

6.1 Visual Tracking of Annealing Process with a Toy Problem

The first experiment shows how the VMA2 method can solve the combinatorial optimization. Consider 3 variables and 1 factor, so that $x_i = (\phi_1 \cdot z_1)\lambda_{1i} + \nu_i$ where all parameters except ϕ_1 are fixed as $\Psi = I$ and $\Delta = I$. The likelihood function in (2) is then $p(X|Z,\Theta) \propto \exp(\phi'_{z_1} S \phi_{z_1}/2)$. Assume that true edge on $z_{31} = 1$, indicating $x_{i3} \leftarrow \lambda_{i1}$, is known, but $z_{11} = 1$ and $z_{21} = 0$ are treated as unknown. Then, with the prior for z_{11} and z_{21} as $\text{logit}(\Pr(z_{11} = 1)) = \text{logit}(\Pr(z_{21} = 1)) = -1.5$, we explored values for ϕ_1 and ω_{g1}, $g = 1,2$, based on an artificial data set drawn from the GFM, so as to refine $\mathcal{G}_T(\Theta,\omega)$ under the factorized $\omega(Z) = \omega(z_{11})\omega(z_{21})$.

We can map the surface $\mathcal{G}_T(\Theta,\omega)$ over $(\omega_{11},\omega_{21})$ when Θ is set at the optimized value $\hat{\Theta}$ for each $(\omega_{11},\omega_{21})$. Figure 2 on the bottom-right corner displays a contour plot of $\mathcal{G}_0(\hat{\Theta},\omega)$. The maximum point lies in one of the four corners corresponding to $\omega_{g1} \in \{0,1\}$ and the global MAP estimate has $\omega_{11} = 1$ and $\omega_{21} = 0$.

Figure 2 also shows a tracking result of the VMA2 search process starting from $T = 2$ and stopping at $T = 0$. The change in $\mathcal{G}_T(\hat{\Theta},\omega)$ and the corresponding maximizing values of $(\omega_{11},\omega_{21})$ can be monitored through the contour plots at selected temperatures. Starting from the initial values, $\omega_{11} \approx 0.5$ and $\omega_{21} \approx 0.5$, at the highest temperature, the successively-generated maximum points gradually come closer to the global optimum ($\omega_{11} = 1$ and $\omega_{21} = 0$) as the annealing process proceeds. At higher temperatures, $\mathcal{G}_T(\hat{\Theta},\omega)$ is unimodal. In the overall search, the tempered criterion begins to become bimodal after the trajectory moves into regions close to the global maximum.

This simple illustrative example highlights the key to success in the search: moving the trajectory of solutions closer to the global maximum in earlier phases of the cooling schedule, before the tempered criterion function exhibits substantial multimodality. Looking ahead, we may be able to raise the power of the annealing search by, for example, using dynamic control of the cooling schedule or more general penalty functions for ω.

6.2 Snapshot of Algorithm with 30 Variables and 4 Factors

In what follows, we will show some simulation studies to provide insights and comparisons. The data sets have $n = 100$ data points drawn from the GFM with $p = 30$ and $k_{\text{true}} = 4$, and with $\Psi = 0.05I$ and $\Delta = \text{diag}(1.5, 1.2, 1.0, 0.8)$. The z_{gj} were independently generated with $\Pr(z_{gj} = 1) = 0.3$, yielding roughly 70% sparsity; then, non-zero elements of Φ where generated as independent standard normal variates, following which Φ_Z was constrained to orthogonality.

To explore sensitivity to the chosen temperature schedule for annealing, experiments were run using three settings:

- (Log-inverse decay) $T_i = 3/\log_2(i+1)$ for $i = 1,\ldots,6999$, and $T_{7000} = 0$
- (Linear decay) $T_i = 3 - 6 \times 10^3 \times (i-1)$ for $i = 1,\ldots,1999$, and $T_{2000} = 0$
- (Power decay) $T_i = 3 \times 0.99^{-(i-1)}$ for $i = 1,\ldots,1999$, and $T_{2000} = 0$

For each, we evaluated the resulting MAP search in terms of comparison with the true model and computational efficiency, in each case using a model with redundant factor dimension $k = 8$.

6.3 Annealing with Fixed Hyper-parameters

First analyses fixed $\zeta_{gj} = c$ and was run repeatedly across some grid points of $c \in [0,5]$. Figure 3 summarizes the evaluation of the receiver operating characteristics (ROC) for the three cooling

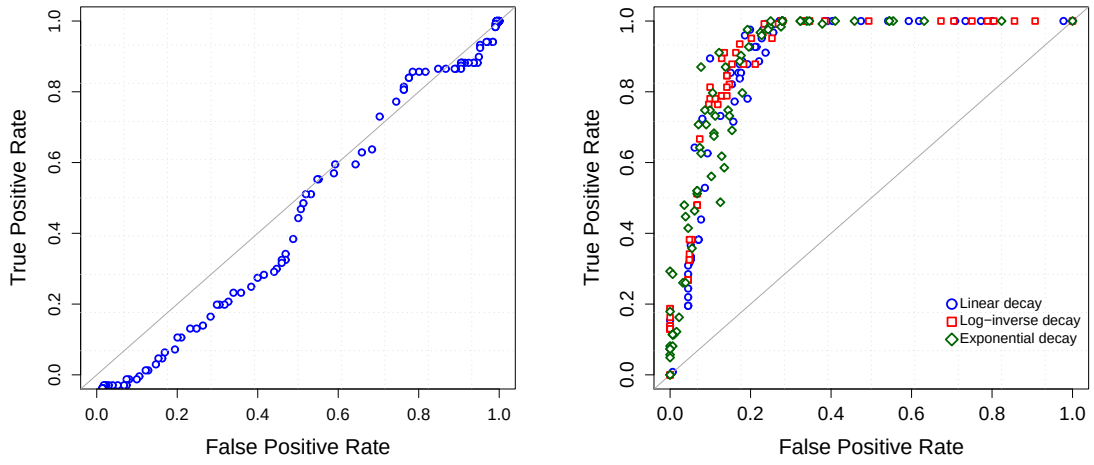

Figure 3: ROC for threshold PCA assuming a known, true $k = 4$ factors (left), compared to VMA2 estimation of the GFM under the three cooling schedules and with $k = 8$ (right). TPR (vertical) and FPR (horizontal) were calculated according to TP/P and FP/N where P and N denote the numbers of non-zero and zero elements in true loadings, TP and FP are the numbers of true positives and false positives, respectively.

schedules. The true positive (TPR) and false positive rates (FPR) were computed based on the correspondences between estimated and true values of the z_{gj}. For comparison, we used standard PCA, extracting the dominant 4 eigenvectors and setting entries below a threshold (in absolute value) to zero; sliding the threshold towards zero gives a range of truncated loadings vectors in the PCA that define the ROC curve for this approach. The resulting ROC curve, shown in the left panel, is very near to the 45° line, comparing very poorly with the annealed GFM; for the latter, each ROC curve indicates rather accurate identification of the sparse structure and the curves differ in small ways only as a function of cooling schedule. The choice of cooling schedule can, however, have a more marked influence on results if initialized at temperatures that are too low.

6.4 Inference on Degrees of Sparseness

A second analysis uses the sparsity prior $p(\zeta_{gj}) = \mathcal{N}_+(\zeta_{gj}|\mu,\sigma)$ with $\mu = 3$ and $\sigma = 6$, and adopts the log-inverse cooling schedule. As shown in the right panel of Figure 4, the analysis realized a reasonable control of FNR (15.4%) and FPR (0%), inducing a slightly less sparse solution than the true structure. The GFM analysis automatically prunes the redundant factors, identifying the true model dimension. Figure 5 displays a snapshot of evolving configuration probabilities ω_{gj} and hyper-parameters ζ_{gj} during the annealing schedule, demonstrating convergence over 2000 steps. At around $T_i \simeq 0.45$, all the configuration probabilities corresponding to the redundant four factors reached to zero.

We further evaluated sensitivity to the choice of cooling schedules; in addition to the previous three cooling schedules, we compared with:

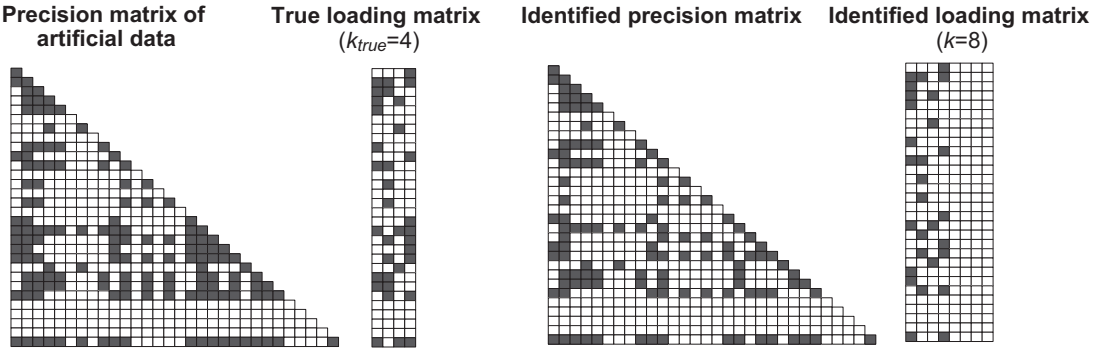

Figure 4: Result of the VMA2 estimation using the log-inverse rate cooling in analysis of synthetic data. (Left) Precision and factor loadings matrix used for generating the synthetic data ($p = 30$, $k_{\text{true}} = 4$). Non-zero elements are colored black. (Right) Estimated precision and factor loadings matrix ($k = 8$); note that the MAP estimate sets the last four loading vectors to zero and so identifies the true number of factors automatically.

- (Log-inverse decay) $T_i = 0.7/\log_2(i+1)$ for $i = 1,\ldots,6999$ and, $T_{7000} = 0$
- (Linear decay) $T_i = 0.7 - 6\times 10^3 \times (i-1)$ for $i = 1,\ldots,1999$ and, $T_{2000} = 0$
- (Power decay) $T_i = 0.7 \times 0.99^{-(i-1)}$ for $i = 1,\ldots,1999$ and, $T_{2000} = 0$

The initial temperatures are reduced from 3 to 0.7. Figure 6 shows the variations of TPR and FPR in the use of the six cooling schedules, evaluated in 20 analyses with replicated synthetic models and data sets. The left and center panels indicate significant dominance of the annealing starting from the higher initial temperatures. Performance in identifying model structure seriously degrades when using a temperature schedule that starts too low, and the sensitivity to schedule is very limited when beginning with reasonably high initial temperatures.

The right panel in Figure 6 shows TPR and FPR for the sparse PCA (SPCA) proposed by Zou et al. (2006), evaluated on the same 20 data sets using the `R` code `spca()` available at CRAN (`http://cran.r-project.org/`). With `spca()`, we can specify the number of nonzero elements (cardinality) in each column of the factor loading matrix. We executed `spca()` after the assignment of the true cardinality as well as the known factor dimension $k_{\text{true}} = 4$. The figure indicates a better performance of GFM annealed with high initial temperature than the sparse PCA, and this is particularly notable in that the GFM analysis uses $k = 8$ and involves no *a priori* knowledge on the degree of sparseness. It is important to see that the conducted comparison is biased since the data were drawn from the GFM with the orthogonal loading matrix where SPCA does not make orthogonality assumptions. In Section 7.1, we provide deeper comparisons among several existing sparse factor analyses based on image processing in hand-written digits recognition.

6.5 Computing Time Questions

Figure 7 shows the CPU times required for the execution of the GFM analyses as above, repeated with increasing dimension $p \in \{100, 200, 300, 500, 700, 1000\}$. The data sets were again generated from GFMs with 4 factors and roughly 70% sparseness. We then performed the VMA2 using a

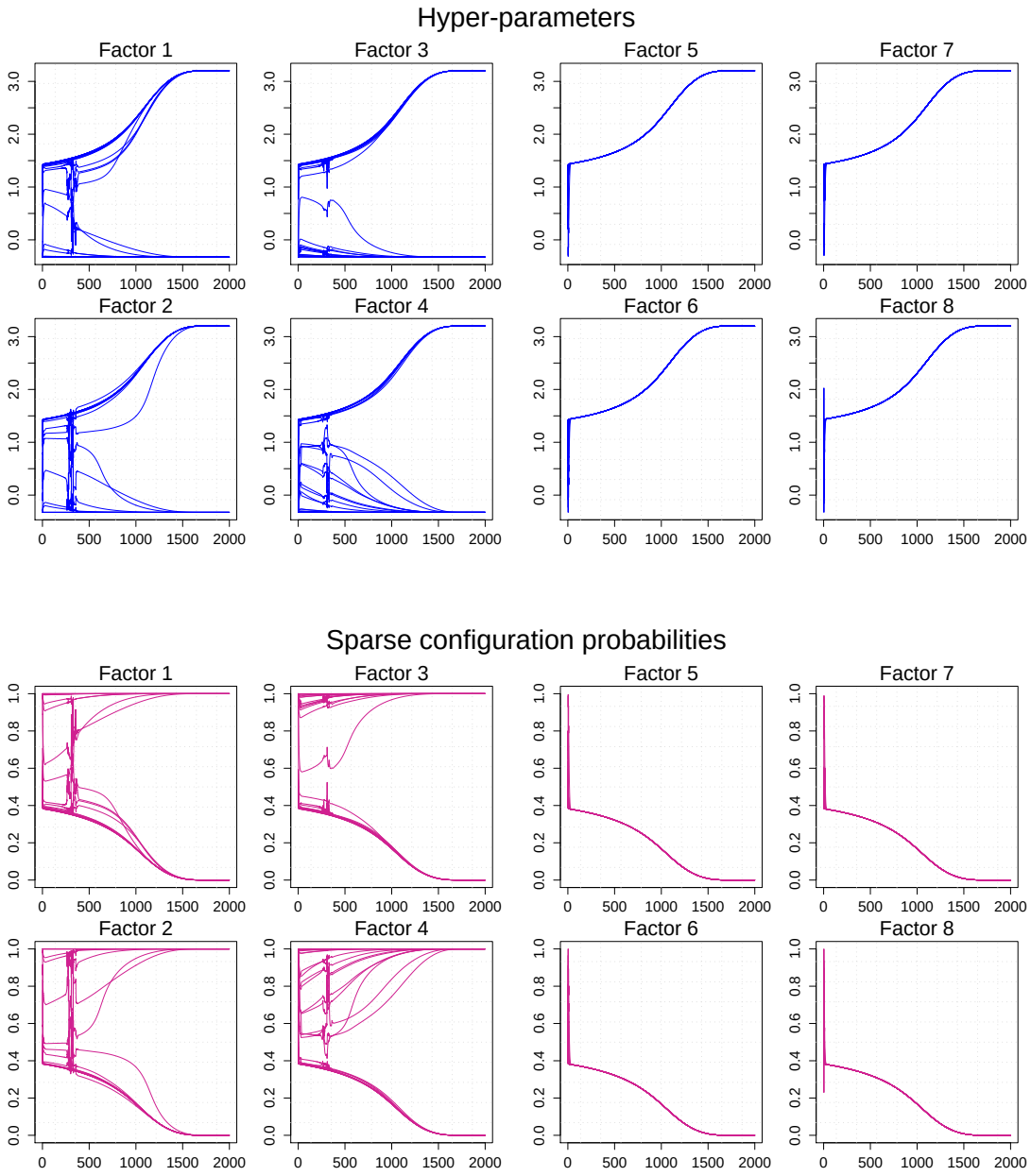

Figure 5: Convergence trajectories of the ζ_{gj} (upper) and ω_{gj} (lower) in analysis of synthetic data over 2000 steps of annealed MAP estimation.

linear decay cooling of length 2000, and using both deterministic and stochastic annealing in a model with $k = 8$. The deterministic algorithm was not used for $p \geq 500$ due to substantial increase in CPU times; this was eased via use of the stochastic search algorithm.

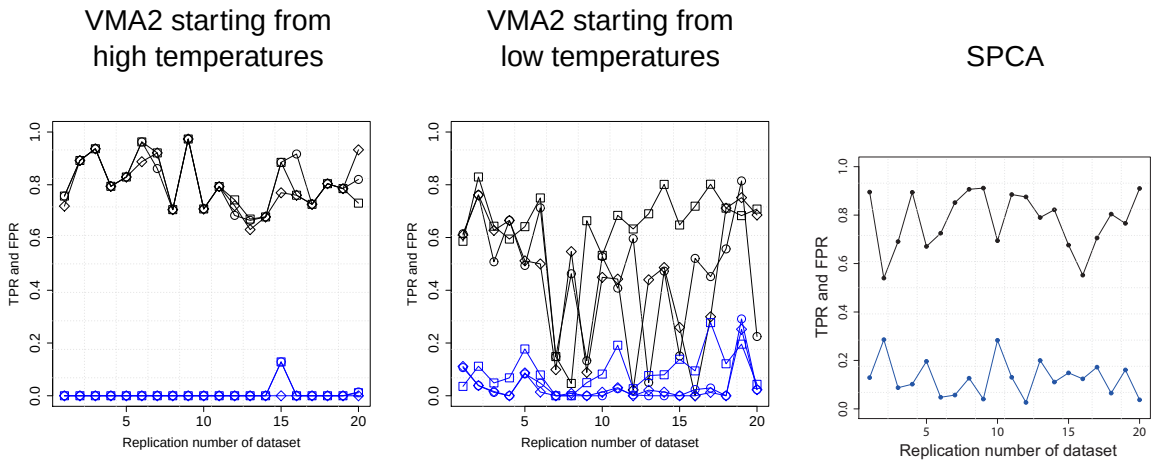

Figure 6: Performance tests on 20 synthetic data sets for different cooling schedules and comparison between the GFM and a sparse PCA (SPCA). For each panel, TPR (black) and FPR (blue) are plotted (vertical axis) against the 20 replicate simulations of artificial data. The results of annealing with the higher and lower initial temperatures are shown in the left and center panels respectively where the rates of cooling with log-inverse, linear and power decays are denoted by box, diamond and circle, respectively. The right panel shows the results of SPCA.

7. Real Data Applications

Experimental results on image analyses of hand-written digits (Section 7.1) and breast cancer gene expression data (Section 7.2) are shown to demonstrate practical relevance of the GFMs in analyses of high dimensional data.

7.1 Application: Hand-written Digit Recognition

We evaluate GFM in pattern recognition analyses of hand-written digit images, and make comparisons to three existing methods; (i) SPCA (Zou et al., 2006), (ii) sparse probabilistic PCA with ARD prior (Archambeau and Bach, 2009), and (iii) MCMC-driven sparse factor analysis (West, 2003; Carvalho et al., 2008). These three methods are all based on models with non-orthogonal sparse loading matrices. The training data set was made from 16×16 gray-scale images of 100 digits (i.e., $n = 100$, $p = 256$) of '3' that were randomly drawn from the postal zip code data available at `http://www-stat.stanford.edu/~tibs/ElemStatLearn/` (Hastie et al., 2001). To evaluate robustness of the four approaches, we added artificial outliers to 15 pixels (features) for about 5% of the original 100 images. Some of the contaminated images are shown in the top-left panels of Figure 8.

For the non-probabilistic method, that is, (i) SPCA, we performed data reconstruction in the standard manner; $x(x_i) = WW'x_i$ with W the matrix of sparse, non-orthogonal loading vectors. In applications of (ii) and (iii) that are inherently driven by probabilistic models, data reconstruction was made via the posterior mean $x(x_i) = W\mathbb{E}[\lambda_i|x_i]$ using obtained sparse loading matrix W. For all

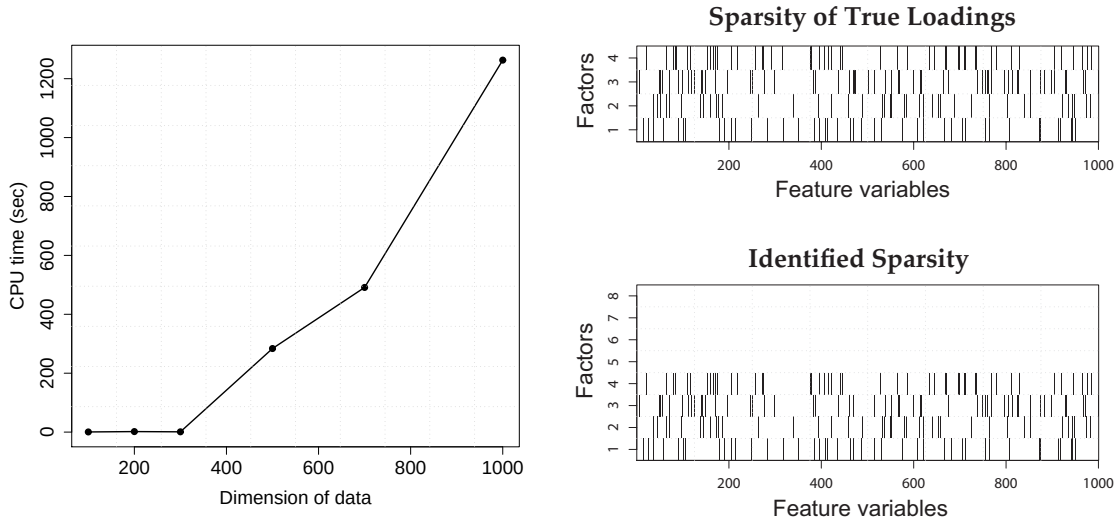

Figure 7: (Left) CPU times (in seconds; Intel(R) Core(TM)2 Duo processor, 2.60Ghz) versus model dimension p for the stochastic VMA2. For the deterministic VMA2, we terminated the tests with the data larger than $p = 300$. Execution times for the deterministic algorithm were approximately 468, 812 and 1100 sec for $p = 100$, 200 and 300. (Right) Identified sparse loadings matrix, displayed as transpose, for the case of $p = 1000$ where the MAP estimation achieved FPR $= 12.0\%$ and FNR $= 18.4\%$.

the methods, setting factor dimensions to $k = 10$, we explored sparse estimates so that the degrees of sparseness become approximately 30% (see Figure 8). For SPCA, we use the same number of non-zero elements in each loading vector as in the estimated GFM. The GFM was estimated using VMA2 with a fixed value for ζ and a linear cooling schedule of length 2000.

A set of 100 test data samples was created from the 100 samples above by adding outliers drawn from a uniform distribution to randomly-chosen pixels with probability 0.2. Performance of the four approaches to data compressions/reconstruction were assessed via mean square error (MSE) between $x(x_i)$s and the true, original test images without the outliers. The right four panels in Figure 8 show some digit images reconstructed by each method with the corresponding original/contaminated test data. The reconstruction errors for the training and test instances are also summarized in the figure. For the results on (ii) and (iii)—the non-orthogonal probabilistic analyses—the reconstructed digits were vaguely-outlined. Such poor reconstructions arise partly from effects of the outliers spread from pixel to pixel along the complete graph defined by non-sparse precision matrix. This empirical result indicates the vulnerability issue of non-restricted sparse factor models in presence of outliers. In the reconstructions of the test instances, the GFM could capture characteristics of original digits with the highest accuracy among the methods. SPCA attained the second highest accuracy in terms of MSE. These observations highlight the substantial merit of using sparse linear mapping in data reconstructions. The GFM and SPCA limit the propagations of outliers within some factor cliques, as most pixel images in the other isolated, non-adjacent factor cliques could be restored clearly.

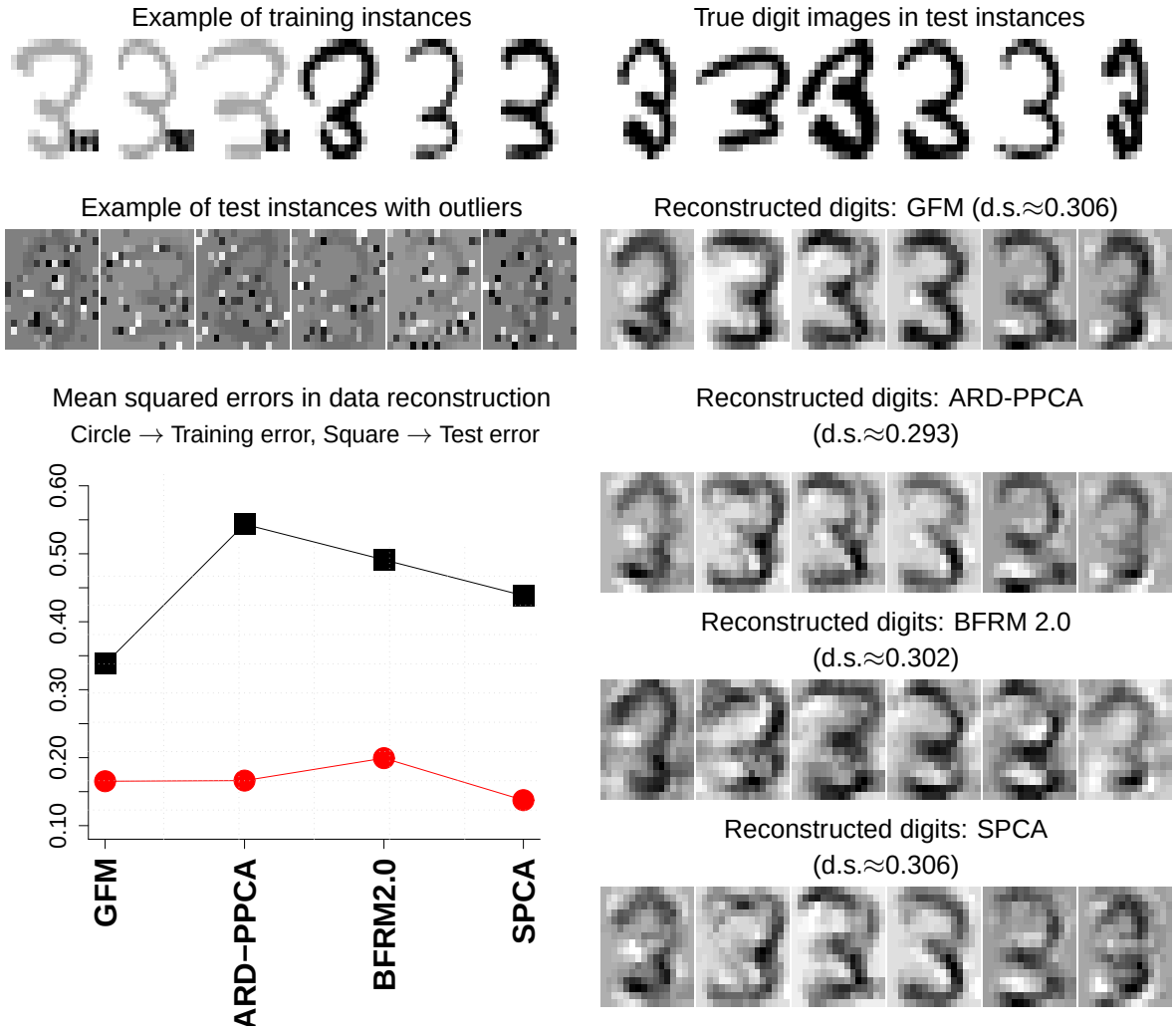

Figure 8: Comparison between GFM and the three alternative methods ((i)-(iii)) in the data reconstruction of outlier hand-written digit images. For the implementation of the sparse probabilistic PCA with ARD prior (ARD-PPCA), we prepared our own `R` function which is available at Supplementary web site. In the application of the MCMC-based sparse factor analysis, we used BFRM 2.0 distributed at http://www.stat.duke.edu/research/software/west/bfrm/. In the four panels on the bottom-right, *d.s.* denotes the degree of sparseness.

7.2 Application: Breast Cancer Gene Expression Study

Latent factor models are being more used in microarray-based gene expression studies in both basic biological and clinical studies, such as cancer genomics. An example in breast cancer gene expression study here further illustrates the practical relevance of GFM structure and adds to comparisons with other approaches. In addition to the summary details below, a much extended discussion of

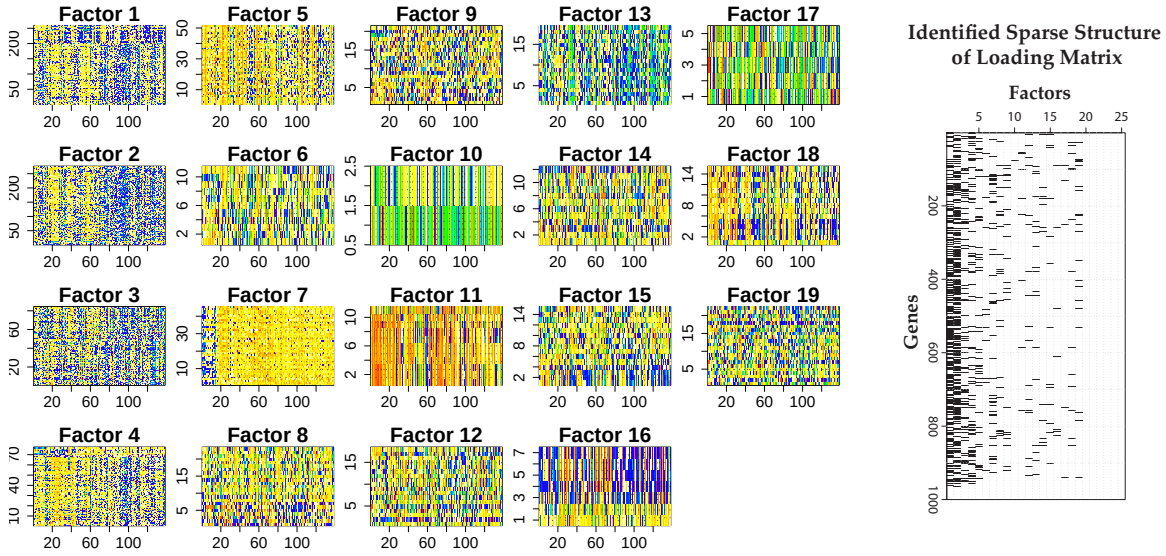

Figure 9: Identified factor probes (left) and sparse structure (right; binary matrix). In each image of the left panel, expression signatures of the probes associated with each factor are depicted across 138 samples (ordered along horizontal axis).

both statistical and biological aspects is available in supporting material at the first author's web site (see link below).

Among the goals of most such studies are identification of multiple factors that may represent underlying activity of biological pathways and provide opportunities for improved predictive models with estimated latent factors for physiological and clinical outcomes (e.g., Carvalho et al., 2008; Chang et al., 2009; Hirose et al., 2008; Lucas et al., 2006, 2009; West, 2003; Yoshida et al., 2004, 2006). Here we discuss an example application of our sparse GFM in analysis of data from previously published breast cancer studies (Cheng et al., 2006; Huang et al., 2003; Pittman et al., 2004; West et al., 2001).

The GFM approach was applied to a sample of gene expression microarray data collected in the CODeX breast cancer genomics study (Carvalho et al., 2008; Pittman et al., 2004; West et al., 2001) at the Sun-Yat Sen Cancer Center, Taipei, during 2001-2004. In addition to summary expression indices from Affymetrix Human Genome U95 arrays, the data set includes immunohistochemistry (IHC) test for key hormonal receptor proteins in clinical prognostics; ERBB2 (Her2) and estrogen (ER). The IHC measures are discrete: ER negative (ER=0), ER positive with low/high-level expression (ER=1 and ER=2), Her2 negative (Her2=0), and Her2 positive with low/high-level (Her2=1 and Her=2). We performed analysis of $p = 996$ genes with the expression levels that, on a log2 (fold change) scale, exceed a median level of 7 and a range of at least 3-fold changes across the tumors. The data set, including the expression data and the IHC hormonal measures, are available on-line as supplementary material.

The annealed estimation of GFM was run with $k = 25$, $\mu = 7$ and $\sigma = 10$. The cooling schedule was prescribed by a linearly-decreasing sequence of 2000 temperatures under which the decay rate and initial temperature were set to 0.006 and 3, respectively. The applied GFM identified 19 factors,

pruning from the model maximum $k = 25$. Heatmaps of gene expression for genes identified in each of the factors appear in Figure 9 with the identified sparse pattern of the loadings matrix.

Evaluation and Annotation of Inferred Factors: To investigate potential biological connections of the factors, we evaluated enrichment of the functional annotations shared by genes in each factor through the Gene Ontology (GO). This exploration revealed associations between some factors and GO biological processes; the complete and detailed results, including tables of the GO enrichment analyses for each factor and detailed biological descriptions, are available from the web site of supporting information.

Factors Related to ER: Figure 10 displays boxplots of fitted values of the factor scores for each sample, plotted across all 19 factors and stratified by levels of each of the clinical ER and Her2 (0/1/2) categories. For each sample i, the posterior mean of the factor vector, namely $\hat{\lambda}_i = (I_k + \Delta)^{-1}\Delta\Phi_Z'\Psi^{-1/2}x_i$, is evaluated at the estimated model, providing the fitted values displayed. We note strong association of ER status to factors 8 (GO: hormone metabolic process), 9 (GO: glucose metabolic process, negative regulation of MAPK activity), 12 (GO: C21-steroid hormone metabolic process), 14 (GO: apoptotic program, positive regulation of caspase activity), 18 (GO: M phase of meiotic cell cycle) and 19 (GO: regulation of Rab protein signal transduction). These clear relationships of ER status to multiple factors with intersecting but also distinct biological pathway annotations is consistent with the known complexity of the broader ER network, as estrogen receptor-induced signaling impacts multiple cellular growth and developmentally related downstream target genes and strongly defines expression factors linked to breast cancer progression.

Her2 Status and Oncogenomic Recombination Hotspot on 17q12: Figure 10 indicates factor 16 as strongly associated with Her2 status $(0,1)$ versus 2. Factor 16 significantly loads on only 7 genes that include STARD3, GRB7 and two probe sets on the locus of ERBB2 (which encodes Her2). This is consistent with earlier gene expression studies that have consistently identified a single expression pattern related to Her2 and a very small number of additional genes, and that have found the "low Her2 positives" level(1) to be generally comparable to negatives. Interestingly, we note that STARD3, GRB7 and ERBB2 are all located on the same chromosomal locus 17q12, which is known as PPP1R1B-STARD3-TCAP-PNMT-PERLD1-ERBB2-MGC14832-GRB7 locus. This locus has been reported in many studies (e.g., Katoh and Katoh, 2004) as an oncogenomic recombination hotspot which is amplified frequently in breast tumor cells, and the purely exploratory (i.e., unsupervised) GFM analysis clearly identifies the "Her2 factor" as strongly reflective of increased expression of genes in this hotspot, consistent with the amplification inducing Her2 positivity.

Comparison to Non-sparse Analysis: Finally, we show a comparison to non-sparse traditional PCA. Supplementary Fig.1 and 2 show the estimated factors (principal components) corresponding to the most dominant 19 eigenvalues, stratified by the levels of ER and Her2. The PCA failed to capture the existing factor relevant to Her2-specific phenotypes in the analysed data. Note that the foregoing sparse analysis identified the Her2-relevant factor only through the 7 non-zero loadings. Indeed, our post-analysis has found that the data set contains very few genes exhibiting significant fold change across the Her2 phenotypes. The non-sparse analysis would capture many irrelevant features through too redundant non-zero loadings. The failure of PCA signifies the importance of sparse modelling in handling high-dimensional data having inherently sparse structure.

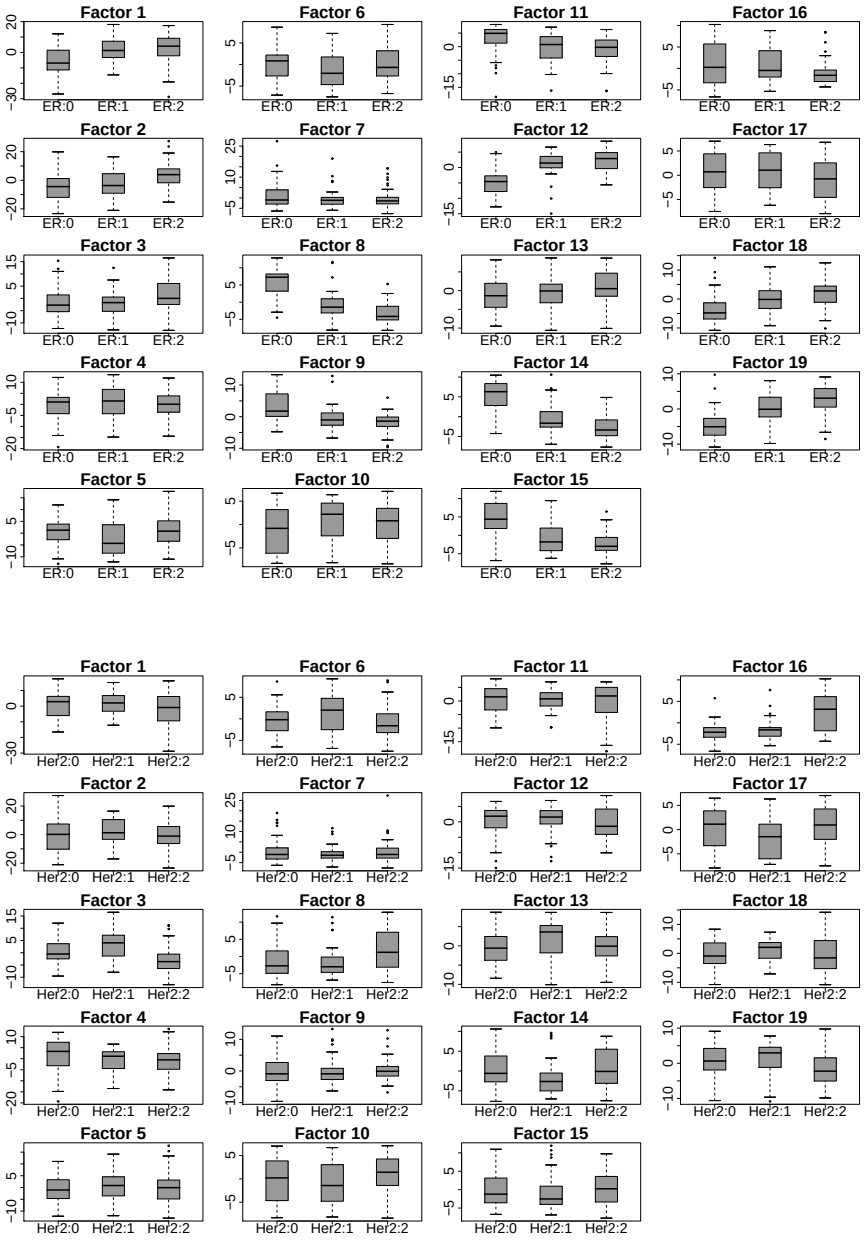

Figure 10: Boxplots of fitted values of breast tumor-specific factor scores, stratified by protein IHC determinations of clinical ER status (upper) and Her2 status (lower) in their 0/1/2 categories.

8. Additional Comments

The novel graphical property of GFMs provides a nice reconciliation of sparse covariance models with sparse precision models—sparse latent factor analysis and graphical models, respectively. Some of the practical benefits of this arise from the ability of GFM to define data reconstructions

exhibiting the same patterns of covariances as the model/data predict, and the potential to induce robustness to outliers relative to non-graphical factor models, whether sparse or not. Some theoretical questions remain about precise conditions under which the sparsity patterns of covariance and precision matrices are guaranteed to agree in general sparse Gaussian factor models other than the GFM form. Additionally, extensions to integrate non-parametric Bayesian model components for factors, following Carvalho et al. (2008), are of clear future interest.

The ability of the VMA2 to aid in the identification of model structure in sparse GFM, and to provide an additional computational strategy and tools to address the inherently challenging combinatorial optimization problem, has been demonstrated in our examples. Scaling to higher dimensional models is enabled by relaxation of the direct deterministic optimization viewpoint, with stochastic search components that promote greater exploration of model space and can speed up search substantially. Nevertheless, moving to higher dimensions will require new, creative computational implementations, such as using distributed computing, that will themselves require novel methodological concepts.

The annealed search methodology evidently will apply in other contexts beyond factor models. At one level, sparse factor models are an instance of problems of variable selection in multivariate regression, in which the regression predictors (feature variables) are themselves unknown (i.e., are the factors). The annealed entropy approach is therefore in principle applicable to problems involving regression model search and uncertainly in general classes of linear or nonlinear multivariate regression with potentially many predictor variables. Beyond this, the same can be said about potential uses in other areas of graphical modelling involving structural inference of directed or undirected graphical models, and also in multivariate time series problems where some of the sparse structure may relate to relationships among variables over time.

We also remark on generalization of the basic form of VMA2 here that might use penalty functions other than the Shannon's entropy used here. The central idea of the VMA2 is the design of a temperature-controlled iterative optimization that converges to the joint posterior distribution of model parameters and sparse structure indicators. The entropy formulation used in our GFM context was inspired by the form of the posterior itself, but similar algorithms—with the same convergent property—could be designed using other forms. This, along with computational efficiency questions and applications in models beyond the sparse GFM framework, and also potential extensions to consider heavy-tailed or Bayesian nonparametric distributions for latent factors and/or residuals (e.g., Carvalho et al., 2008), are open areas for future research.

Acknowledgments

The authors are grateful to the Action Editor and three anonymous referees for their detailed and most constructive comments on the original version of this paper. Elements of the research reported here were developed while Ryo Yoshida was visiting SAMSI and Duke University during 2008-09. Aspects of the research of Mike West were partially supported by grants from the U.S. National Science Foundation (DMS-0342172) and National Institutes of Health (NCI U54-CA-112952). The research of Ryo Yoshida was partially supported by the Japan Science and Technology Agency (JST) under the Core Research for Evolutional Science and Technology (CREST) program. Any opinions, findings and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the NSF or NIH.

Appendix A.

We present a proof of Proposition 1 and a derivation of optimization over Φ.

A.1 Proof of Proposition 1

Replace the objective function of (5) by multiplying by inverse temperature $1/T$:

$$\frac{1}{T}\mathcal{G}_T(\Theta,\omega) = \sum_{Z\in\mathscr{Z}} \omega(Z)\log p(X,Z,\Theta|\zeta)^{1/T} - \sum_{Z\in\mathscr{Z}} \omega(Z)\log\omega(Z).$$

An upper-bound of this modified criterion is derived as follows:

$$\begin{aligned}
\frac{1}{T}\mathcal{G}_T(\Theta,\omega) &= \sum_{Z\in\mathscr{Z}} \omega(Z)\log\frac{p(Z|X,\Theta,\zeta)^{1/T}p(X,\Theta|\zeta)^{1/T}}{\omega(Z)} \\
&= \sum_{Z\in\mathscr{Z}} \omega(Z)\log\frac{p(Z|X,\Theta,\zeta)^{1/T}}{\omega(Z)\sum_{Z'\in\mathscr{Z}} p(Z'|X,\Theta,\zeta)^{1/T}} + K_0 \\
&\leq K_0.
\end{aligned}$$

In the second equality, the terms irrelevant to $\omega(Z)$ are included in $K_0 = \log p(X,\Theta|\zeta)^{1/T} + \log\sum_{Z'\in\mathscr{Z}} p(Z'|X,\Theta,\zeta)^{1/T}$. The first term in the second line is the negative of the Kullback-Leibler divergence between $\omega(Z)$ and the normalized tempered posterior distribution. The lower-bound of the Kullback-Leibler divergence is attained if and only if

$$\omega(Z) = \frac{p(Z|X,\Theta,\zeta)^{1/T}}{\sum_{Z'\in\mathscr{Z}} p(Z'|X,\Theta,\zeta)^{1/T}},$$

as required.

A.2 Derivation: Optimization over Φ

Let ρ_j, $j\in\{1,\ldots,k\}$ be the Lagrange multipliers to ensure the restrictions in (10). We now write down the Lagrange function:

$$\phi_j'\mathbb{E}_\omega[S(z_j,\Psi)]\phi_j - \rho_j(\|\phi_j\|^2-1) - \sum_{m\neq j}\rho_m\phi_m'\phi_j. \tag{13}$$

Differentiation of (13) with respect to ϕ_j yields

$$\mathbb{E}_\omega[S(z_j,\Psi)]\phi_j - \rho_j\phi_j - \sum_{m\neq j}\rho_m\phi_m = 0. \tag{14}$$

In order to solve this equation, the first step to be addressed is to find the closed form solution for the vector of the Lagrange multipliers, $\rho_{(-j)} = \{\rho_m\}_{m\neq j}\in\mathbb{R}^{k-1}$. Multiplying (14) by each ϕ_m', $m\neq j$, from the left, we have the $k-1$ equations as follows:

$$\phi_m'\mathbb{E}_\omega[S(z_j,\Psi)]\phi_j - \sum_{m\neq j}\rho_m\phi_m'\phi_j = 0 \quad \text{for } m \text{ s.t. } m\neq j.$$

This yields the matrix representation

$$\Phi'_{(-j)}\mathbb{E}_\omega[S(z_j,\Psi)]\phi_j - \Phi'_{(-j)}\Phi_{(-j)}\rho_{(-j)} = 0,$$

which in turn leads to the solution for $\rho_{(-j)}$ as

$$\rho_{(-j)} = (\Phi'_{(-j)}\Phi_{(-j)})^{-1}\Phi'_{(-j)}\mathbb{E}_\omega[S(z_j,\Psi)]\phi_j.$$

Substituting this into the original Equation (14) yields the eigenvalue equation

$$N_j\mathbb{E}_\omega[S(z_j,\Psi)]\phi_j - \rho_j\phi_j = 0 \quad \text{with } N_j = I - \Phi_{(-j)}\Phi'_{(-j)}. \tag{15}$$

Now consider the alternative, symmetrized eigenvalue equation

$$N_j\mathbb{E}_\omega[S(z_j,\Psi)]N_j\varphi_j - \rho_j\varphi_j = 0. \tag{16}$$

Since N_j is idempotent, left-multiplication of (16) by N_j yields

$$N_j\mathbb{E}_\omega[S(z_j,\Psi)]N_j\varphi_j - \rho_j N_j\varphi_j = 0.$$

which is equivalent to the required Equation (15) when $\phi_j = N_j\varphi_j$.

References

T.W. Anderson. *An Introduction to Multivariate Statistical Analysis, 3rd ed.* Wiley-Interscience; New Jersey, 2003.

C. Archambeau and F. Bach. Sparse probabilistic projections. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 73–80, Cambridge, MA, 2009. MIT Press.

H. Attias. Inferring parameters and structure of latent variable models by variational Bayes. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI99)*, pages 21–30, 1999.

C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society (Series B)*, 61:611–622, 1999.

C.M. Bishop. *Pattern Recognition and Machine Learning, 1st ed.* Springer: Singapore, 2006.

C.M. Carvalho and M. West. Dynamic matrix-variate graphical models. *Bayesian Analysis*, 2: 69–98, 2007.

C.M. Carvalho, J.E. Lucas, Q. Wang, J.T. Chang, J.R. Nevins, and M. West. High-dimensional sparse factor modelling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103:1438–1456, 2008.

J. Chang, C.M. Carvalho, S. Mori, A. Bild, Q. Wang, M. West, and J.R. Nevins. Decomposing cellular signaling pathways into functional units: A genomic strategy. *Molecular Cell*, 34:104–114, 2009.

S.H. Cheng, M. West C.F.Horng, E. Huang, J. Pittman, H. Dressman M.H. Tsou, C.M. Chen, S.Y. Tsail, J.J. Jian, J.R. Nevins M.C. Liu, and A.T. Huang. Genomic prediction of loco-regional recurrence following mastectomy in breast cancer. *Journal of Clinical Oncology*, 24:4594–4602, 2006.

A. d'Aspremont, L. El Ghaoui, M.I. Jordan, and G.R.G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007.

A. Dobra, B. Jones, C. Hans, J.R. Nevins, and M. West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196–212, 2004.

T.L. Griffiths and Z Ghahramani. Infinite latent feature models and the indian buffet process. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 475–482, Cambridge, MA, 2006. MIT Press.

Y. Guan and J Dy. Sparse probabilistic principal component analysis. *Proceedings of AISTATS 2009*, 5:185–192, 2009.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Newv York: Springer, 2001.

O. Hirose, R. Yoshida, S. Imoto, R. Yamaguchi, T. Higuchi, D. S. Charnock-Jones, C. Print, and S. Miyano. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics*, 24(7):932–942, 2008.

E. Huang, S Chen, H. K. Dressman, J. Pittman, M. H. Tsou, C. F Horng, A. Bild, E. S. Iversen, M. Liao, C. M. Chen, M. West, J. R. Nevins, and A. T. Huang. Gene expression predictors of breast cancer outcomes. *The Lancet*, 361:1590–1596, 2003.

I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 112(3):531–547, 2003.

B. Jones, A. Dobra, C.M. Carvalho, C. Hans, C. Carter, and M. West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20:388–400, 2005.

M.I. Jordan, editor. *Learning in Graphical Models*. Cambridge MA: MIT Press, 1999.

M.I. Jordan. Graphical models. *Statistical Science*, 19:140–155, 2004.

M. Katoh and M. Katoh. Evolutionary recombination hotspot around GSDML-GSDM locus is closely linked to the oncogenomic recombination hotspot around the PPP1R1B-ERBB2-GRB7 amplicon. *International Journal of Oncology*, 24:757–63, 2004.

J.E. Lucas, C.M. Carvalho, Q. Wang, A.H. Bild, J.R. Nevins, and M. West. Sparse statistical modelling in gene expression genomics. In P. Müller, K.A. Do, and M. Vannucci, editors, *Bayesian Inference for Gene Expression and Proteomics*, pages 155–176. Cambridge University Press, 2006.

J.E. Lucas, C.M. Carvalho, J-T.A. Chi, and M. West. Cross-study projections of genomic biomarkers: An evaluation in cancer genomics. *PLoS ONE*, 4(2):e4523, 2009.

D.J.C. Mackay. Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6:469–505, 1995.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML2009)*, pages 689–696, New York, NY, USA, 2009. ACM.

J. Pittman, E. Huang, H. Dressman, C. F. Horng, S. H. Cheng, M. H. Tsou, C. M. Chen, A. Bild, E. S. Iversen, A. T. Huang, J. R. Nevins, and M. West. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences*, 101:8431–8436, 2004.

P. Rai and H. Daumé. The infinite hierarchical factor regression model. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1321–1328, Cambridge, MA, 2009. MIT Press.

N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural Networks*, 11:271–282, 1998.

M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

Q. Wang, C.M. Carvalho, J.E. Lucas, and M. West. BFRM: Bayesian factor regression modelling. *Bulletin of the International Society for Bayesian Analysis*, 14(2):4–5, 2007.

M. West. Bayesian factor regression models in the "large p, small n" paradigm. In J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West, editors, *Bayesian Statistics 7*, pages 723–732. Oxford University Press, 2003.

M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, H. Zuzan R. Spang, J.R. Marks, and J.R. Nevins. Predicting the clinical status of human breast cancer utilizing gene expression profiles. *Proceedings of the National Academy of Sciences*, 98:11462–11467, 2001.

R. Yoshida, T. Higuchi, and S. Imoto. A mixed factors model for dimension reduction and extraction of a group structure in gene expression data. *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, pages 161–172, 2004.

R. Yoshida, T. Higuchi, S. Imoto, and S. Miyano. Arraycluster: an analytic tool for clustering, data visualization and module finder on gene expression profiles. *Bioinformatics*, 22(12):1538–1539, 2006.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.