Discriminative Mixtures of Sparse Latent Fields for Risk Management

Felix V. Agakov
Pharmatics Ltd, Edinburgh, UK
felix@pharmaticsltd.com

Peter Orchard
University of Edinburgh
peter.orchard@gmail.com

Amos Storkey
University of Edinburgh
amos@storkey.org

Abstract

We describe a simple and efficient approach to learning structures of sparse high-dimensional latent variable models. Standard algorithms either learn structures of specific predefined forms, or estimate sparse graphs in the data space ignoring the possibility of the latent variables. In contrast, our method learns rich dependencies and allows for latent variables that may confound the relations between the observations. We extend the model to conditional mixtures with side information and non-Gaussian marginal distributions of the observations. We then show that our model may be used for learning sparse latent variable structures corresponding to multiple unknown states, and for uncovering features useful for explaining and predicting structural changes. We apply the model to real-world financial data with heavy-tailed marginals covering the low- and high- market volatility periods of 2005-2011. We show that our method tends to give rise to significantly higher likelihoods of test data than standard network learning methods exploiting the sparsity assumption. We also demonstrate that our approach may be practical for financial stress-testing and visualization of dependencies between financial instruments.

1 Introduction

Finding structure in data is one of the most fundamental problems of data analysis and machine learning. In this paper we note that learning and understanding relationships between variables may be viewed as probabilistic inference under *sparsity constraints* on the structure of an underlying latent variable model. A statement that data is "structured" often implies existence of a sparsely connected graphical model (such as a tree or a bipartite graph) that gives rise to the observations, whereas "unstructured" data is often too complex to be accurately represented by a sparse network. The ability to efficiently discover sparse representations of data is fundamental for knowledge representation, interpretability of inferences, computational tractability, and generalization, with applications ranging from systems biology to social marketing and finance.

Standard methods for learning network structures are based on expensive and difficult-to-analyze combinatorial search, where a huge space of possible networks is traversed heuristically to find networks scoring highly by some measure, and/or satisfying conditional independence constraints. The vast majority of such approaches cannot be easily extended to handle latent variables, or may only be used for limited classes of low-dimensional models (e.g. [8, 19, 16, 36, 11, 29] and references therein). Recently [21] described an extension that selects the best fitting model from several candidate structural forms. Other methods may only be used to learn structures of specific predefined forms, with hard constraints on the number of node parents and their cardinality in a directed network [41, 18]. While these approaches are justified when it is known *a priori* that the structure belongs to a standard class of models, they may not be appropriate in a more general real-world setting with richer underlying models. In some cases an explicit constraint on the structural form of a model can make the inference problem more complex than necessary, and so not scalable to higher dimensions.

Recently there has been much work on learning sparse structures of fully observed Gaussians (e.g. [5, 14, 27, 24]). In contrast to combinatorial search approaches relying on greedy heuristics, these more specialized methods view structure learning as the problem of *continuous* optimization. The only constraint specified explicitly is sparsity of the underlying Gaussian in the data space $p(\mathrm{y})$, which is enforced by imposing sparseness-inducing penalties on the elements of the precision matrix. A simple approach to the resulting optimization problem is the graphical LASSO (GLASSO) of [14], that has been successfully applied to estimating sparse inverse covariances for thousands of vari-

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume 22 of JMLR: W&CP 22.

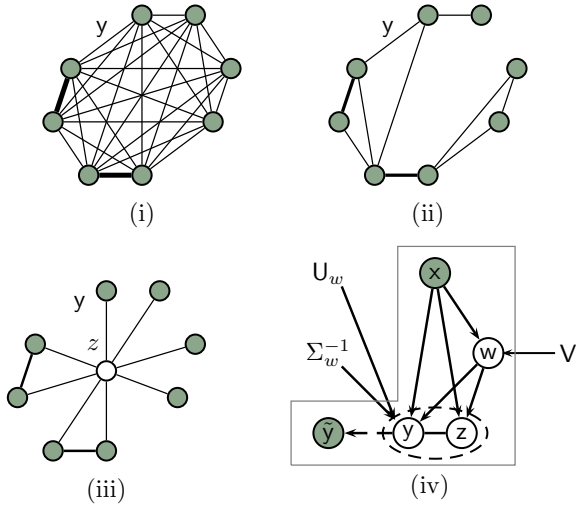

Figure 1: (i) Fully-observed model with a densely-structured marginal $p(\mathsf{y})$ (shaded nodes correspond to the observed variables y; thicker lines indicate stronger pairwise potentials). (ii) Structure of a fully observed model with a sparsity constraint on $p(\mathsf{y})$. Sparsity in the data space may give rise to extraneous conditional independences and results in a pruning of potentially important regularities. (iii) Sparse latent field $p(\mathsf{y}, \mathsf{z})$ with a dense marginal $p(\mathsf{y})$ (latent variables z are represented by transparent nodes). Note that in contrast to latent factor models, the conditional $p(\mathsf{y}|\mathsf{z})$ may contain residual couplings. (iv) Discriminative mixture of latent field models. Each expert $p(\mathsf{y}, \mathsf{z}|w, \mathsf{x})$ has a sparse structure in the augmented space $\{\mathsf{y}, \mathsf{z}\}$, which is enforced by imposing a sparsity-inducing prior on the undirected structure specified by $\mathbf{\Sigma}_w^{-1}$. Side information x is taken into account by explicitly parameterizing the gating distribution $p(w|\mathsf{x}; \mathsf{V})$ and experts $p(\mathsf{y}, \mathsf{z}|\mathsf{x}, w; \mathsf{U}_w, \mathbf{\Sigma}_w^{-1})$. The mapping $\mathsf{y} \rightarrow \tilde{\mathsf{y}}$ is one-to-one. In the case of discriminative mixtures of sparse latent Gaussians, $\tilde{\mathsf{y}} \equiv \mathsf{y}$. In the case of copulas, y is a Gaussianized representation of a non-Gaussian observation $\tilde{\mathsf{y}}$.

ables; other efficient optimization approaches are reviewed and compared in e.g. [15]. The key assumption of these methods is that the data is complete, i.e. there are no missing observations or latent factors influencing the relations between the modeled variables.

This work is motivated by several simple observations about structure in real-world data. First, we observe that the notion of structuredness of data is tightly linked to the sparsity of the underlying data representations in the *augmented space* of the visible and hidden variables $\{\mathsf{y}, \mathsf{z}\}$, rather than sparsity in the data space $\{\mathsf{y}\}$ alone. Here y and z are the observations and latent variables respectively. In situations where some variables may be hidden or missing (which is common in a broad range of real-world applications ranging from finance to systems biology), setting explicit sparsity constraints on the marginals $p(\mathsf{y})$ may lead to pruning of important regularities and result in potentially misleading representations of underlying structures (see Figure 1 (i)–(iii)). Note that sparseness of the joint structure in the augmented space is a standard assumption of many commonly used latent variable models, including latent trees, hidden Markov models, restricted Boltzmann machines, etc., all of which may give rise to potentially dense marginals $p(\mathsf{y})$. While it is commonly assumed that sparse structures are heuristically fixed *a priori*, we are interested in inferring them from data, with the only constraint being sparsity of $p(\mathsf{y}, \mathsf{z})$.

The second important observation is that in many real-life applications, structural dependencies between the variables are hardly ever homogeneous for all data samples and may often depend on poorly understood latent states. For example, some subgroups of cancer patients may be more susceptible to chemotherapy than others, which may be manifested through different structures of proteomic networks. In the financial risk management problem, dependencies between returns of portfolio composites may vary based on market conditions. It is intuitive that a single smooth model may be too coarse to adequately represent finer structure.

The third observation is linked to the fact that inference of the unknown states that may affect structural dependencies may often be facilitated by side (concomitant) information or other explanatory variables. Experts in a subject area are often able to construct very high-dimensional vectors of features which, according to their knowledge of the field, may help to predict structural changes. Also, fund managers may need to be able to *explain* which few macroeconomic or market indicators, types of news announcements, etc. are most predictive of new market regimes, so that they can better motivate changes in their management strategies to their customers. This ability to identify the few important features most strongly affecting the resulting structures is particularly important in *decision support* applications, where the final managerial decisions are left to humans.

The fourth, but rather obvious, observation is that a multivariate Gaussian will rarely be a good model of high-dimensional real-world data. This is the case, for example, in finance [12].

The common methods for high-dimensional sparse structure learning either ignore the four practical observations discussed above - as is the case for fully observed sparse Gaussians MRFs - or address only one of them in a manner that is not easily extensible [28, 6]. Motivated by real-world problems, we extend existing approaches to describe a sparse discriminative mixture of sparse latent Gaussian fields. Our model can be used to learn multiple interpretable latent variable structures, each corresponding to a unique unknown state, and to identify explanatory features useful for predicting structural changes. By extending the

model to a conditional mixture with the components defined by sparse latent copulas (Figure 1 (iv)), we also show that our approach is significantly more accurate than existing methods when the observations are non-Gaussian.

We demonstrate our approach by applying it to the study of the influence of market indicators on relationships between composites of the FTSE100 index of the largest companies listed on the London Stock Exchange. Our experiments cover the high- and low- market volatility period of 2005-2011. We empirically show that our method results in significantly higher test likelihoods than the existing structure learning models exploiting the sparsity assumption. We also show that our model may be attractive for visualizing dependencies between portfolio composites, and demonstrate that it may be practical for financial stress-testing without having to rely on heuristics or expensive expertise.

2 Methodology

Sparsity of models is important for knowledge discovery, generalization, and interpretability of relations between the variables. However, the common assumptions of sparse Gaussian MRFs that (i) the data is fully observed, (ii) the underlying model is sparse in the *data space*, (iii) the dependencies are accurately modeled by the (inverse) covariance of the Gaussian have severe practical limitations. Here we will describe the sparse latent Gaussian field model and its extensions to conditional mixtures of experts that overcome these limitations.

2.1 Graphical LASSO

It is well known that the structure of a Gaussian graphical model $p(\mathsf{y}) \sim \mathcal{N}(., \mathbf{\Sigma})$ is defined by its inverse covariance (precision) matrix $\mathbf{\Sigma}^{-1}$. Variables y_i and y_j ($i \neq j$) are conditionally independent given all the other variables if and only if $(\mathbf{\Sigma}^{-1})_{ij} = 0$ (see e.g. [9], [25]). That is, simpler and more interpretable Gaussian models are represented by precision matrices with greater sparsity. Recently [5] noted that a sparse estimate of the precision $\mathsf{X}_{\mathsf{yy}} \stackrel{\text{def}}{=} \hat{\mathbf{\Sigma}}_{\mathsf{yy}}^{-1} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ may be obtained by optimizing the regularized log-likelihood

$$\max_{\mathsf{X}_{\mathsf{yy}} \succ 0} \left[\log\det \mathsf{X}_{\mathsf{yy}} - \mathrm{tr}\{\mathsf{S}_{\mathsf{yy}}\mathsf{X}_{\mathsf{yy}}\} - \gamma\|\mathsf{X}_{\mathsf{yy}}\|_1\right] \qquad (1)$$

where S_{yy} is the empirical covariance, γ is the regularization penalty, and $\|\mathsf{M}\|_1 = \sum_{ij}|M_{ij}|$. This is equivalent to *maximum-a-posteriori* (MAP) estimation with Laplace priors on the elements of X_{yy} [30]. [5] further shows that (1) is convex, and suggests a block-wise interior-point procedure for optimizing the equivalent dual problem that is guaranteed to converge to a globally optimal positive-definite solution for X_{yy} as long as the initialization $\mathsf{X}_{\mathsf{yy}}{}^{(0)}$ is also positive-definite. Related approaches discussed by [31], [40], and [14] vary mainly in the optimization procedures used. The approach of [14], the graphical LASSO (GLASSO), is based on a fast coordinate descent algorithm, and until very recently was thought to be computationally efficient [15]. In this paper, irrespective of the details of optimization, we will refer to any of the methods for optimizing (1) as graphical LASSO.

2.2 Sparse Latent Inverse Covariance Estimation

In our latent variable extension of GLASSO, we are interested in inferring a sparsely structured X of a Gaussian $p(\mathsf{y}, \mathsf{z}; \mathsf{X})$, where $\mathsf{X} \stackrel{\text{def}}{=} [\mathsf{X}_{\mathsf{yy}}\ \mathsf{X}_{\mathsf{yz}}; \mathsf{X}_{\mathsf{zy}}\ \mathsf{X}_{\mathsf{zz}}]$ is the joint precision. By imposing a sparsity-inducing prior on the structure $p(\mathsf{X}) \propto \exp\{-\gamma\|\mathsf{X}\|_1\}\mathbb{I}(\mathsf{X} \succ 0)$ and considering the MAP approximation of Bayesian inference, the task reduces to optimizing

$$\max_{\mathsf{X}\succ 0} \left[\log\det \mathbf{\Sigma}_{\mathsf{yy}}^{-1} - \mathrm{tr}\{\mathsf{S}_{\mathsf{yy}}\mathbf{\Sigma}_{\mathsf{yy}}{}^{-1}\} - \gamma\|\mathsf{X}\|_1\right] \qquad (2)$$

where $\mathbf{\Sigma}_{\mathsf{yy}}{}^{-1} \stackrel{\text{def}}{=} \mathsf{X}_{\mathsf{yy}} - \mathsf{X}_{\mathsf{yz}}\mathsf{X}_{\mathsf{zz}}{}^{-1}\mathsf{X}_{\mathsf{zy}}$. Since the term $\mathsf{X}_{\mathsf{yz}}\mathsf{X}_{\mathsf{zz}}{}^{-1}\mathsf{X}_{\mathsf{zy}}$ in (2) is generally dense, it is clear that optimization of (2) may result in dense marginals $p(\mathsf{y})$; however, the regularization term $\|\mathsf{X}\|_1$ encourages sparsity of the joint model $p(\mathsf{y}, \mathsf{z})$. In contrast to the GLASSO objective (1), optimization problem (2) is no longer convex in the blocks of X. It is also easy to see that X_{zz} needs to be constrained in order to avoid invariance of the solutions under simultaneous rescaling of X_{zz} and X_{zy}, which may be achieved by penalizing matrices X with small determinants, or explicitly constraining the trace or diagonal elements of X_{zz}. Such constraints are analogous to fixing the variances of the latent factors in factor analysis or random effect models. It is important to note that in contrast to factor analysis, X_{yy} is not necessarily diagonal, i.e. the conditional $p(\mathsf{y}|\mathsf{z})$ is generally not factorized in y. Also, due to the choice of sparse prior on X, the hidden-visible block X_{yz} is generally not rotation-invariant.

To optimize (2), we take the route of a simple and computationally efficient (structural) EM approach for the MAP criterion. In our case the regularized complete-data log-likelihood is given by

$$\mathcal{Q}(\mathsf{X}; \mathsf{X}^{(t-1)}) = \log\det \mathsf{X} - \mathrm{tr}\{\mathsf{S}^{(t-1)}\mathsf{X}\} - \gamma\|\mathsf{X}\|_1 \qquad (3)$$

where $\mathsf{X} \in \mathbb{R}^{(|\mathsf{y}|+|\mathsf{z}|)\times(|\mathsf{y}|+|\mathsf{z}|)}$ is the precision in $\{\mathsf{y}, \mathsf{z}\}$, and $\mathsf{S}^{(t-1)} \stackrel{\text{def}}{=} \mathrm{cov}(\mathrm{vec}\{\mathsf{y}, \mathsf{z}\}|\mathsf{X}^{(t-1)})$ is the estimate of the joint covariance conditioned on the previous structure and averaged over the empirical distribution. The latter is computed in the E-step. The M-step consists of finding the positive semi-definite X that optimizes (3). Note that the objective (3) takes the same form as the graphical lasso objective (1).

From (3) it is straightforward that the M-step of the algorithm reduces to performing an iteration of GLASSO in $\{\mathsf{y}, \mathsf{z}\}$, while the E-step evaluates the moments of the joint

Algorithm 1 Sparse Latent Inverse Covariance Estimation (SLICE)

Initialize $\mathsf{X}^{(0)}$ so that $\mathsf{X}^{(0)} \succ 0$
for $t = 1 : T$ **do**
E-step:
$\mathsf{S}^{(t-1)} \leftarrow \begin{pmatrix} \mathsf{S}_{yy} & \mathsf{S}_{yy}\mathsf{A} \\ \mathsf{A}^T\mathsf{S}_{yy} & \mathsf{X}_{zz}{}^{(t-1)} + \mathsf{A}^T\mathsf{S}_{yy}\mathsf{A} \end{pmatrix}$
where
$\mathsf{A} \stackrel{\text{def}}{=} -\mathsf{X}_{yz}{}^{(t-1)}(\mathsf{X}_{zz}{}^{(t-1)})^{-1}$
M-step:
$\mathsf{X}^{(t)} \leftarrow \arg\max_{\mathsf{X}} \left[\log\det\mathsf{X} - \mathrm{tr}\{\mathsf{S}^{(t-1)}\mathsf{X}\} - \gamma\|\mathsf{X}\|_1\right]$
such that $\mathsf{X} \succ \mathbf{0}$ and $(\mathsf{X}_{zz})_{ii} = \text{const}$
end for

Gaussian with the previous structure. We call this procedure for fitting sparse Gaussian latent fields *SLICE* (Sparse Latent Inverse Covariance Estimation). It is shown in Algorithm 1. The M-step may be based on any of the constrained GLASSO-like optimization procedures [38, 35, 39]. For example, the LogdetPPA method of [39] is suitable for this purpose: it is designed to solve problems of the form

$$\min_{\mathsf{X}} \left[\mathrm{tr}\{\mathsf{SX}\} - \mu\log\det\mathsf{X} : A(\mathsf{X}) = \mathsf{b}, \mathsf{X} \succ 0\right] , \quad (4)$$

where A is a linear map and μ is a constant. Another option is the more general-purpose semidefinite programming package SDPT3 [38]. We found that SDPT3 is faster than LogdetPPA on low-dimensional problems (up to around 40 variables or so), but LogdetPPA scales better to higher dimensions. So we use SDPT3 on smaller problems and LogdetPPA if the problem is larger.

2.3 Discriminative Mixtures of Sparse Gaussians and Conditional Sparse Latent Fields (MSLICE)

SLICE is easily extended to the case of discriminative mixtures

$$p(\mathsf{y},\mathsf{z}|\mathsf{x}) = \sum_{w=1}^{K} p(\mathsf{y},\mathsf{z}|\mathsf{x},w)p(w|\mathsf{x}), \quad (5)$$

where w is a mixture component, $p(w|\mathsf{x})$ is the gating distribution, and $p(\mathsf{y},\mathsf{z}|\mathsf{x},w)$ is the w^{th} expert. Using the standard notation for mixtures of experts [20], we define vectors of mixture indicators $\mathsf{w}^{(i)} \in \{0,1\}^K$ (where $w_j^{(i)} = 1$ iff $\mathsf{x}^{(i)}$ belongs to cluster j) and x is the side information. Figure 1 (iv) shows the graphical model. We considered several parameterizations of the experts $p(\mathsf{y}|\mathsf{w},\mathsf{x})$ and the gating distribution $p(\mathsf{w}|\mathsf{x})$. One such choice is to set

$$\pi_j(\mathsf{x}^{(i)}) \stackrel{\text{def}}{=} p(w_j^{(i)} = 1|\mathsf{x}^{(i)};\mathsf{v}_j) \propto \exp\{-f(\mathsf{v}_j;\mathsf{x}^{(i)})\} \quad (6)$$

$$p(\mathsf{y},\mathsf{z}|\mathsf{x}^{(i)}, w_j^{(i)} = 1;\mathsf{U}_j) \sim \mathcal{N}(\mathsf{U}_j\mathsf{x}^{(i)};\mathsf{X}_j^{-1}) \quad (7)$$

where $\mathsf{U}_j \in \mathbb{R}^{(|\mathsf{y}|+|\mathsf{z}|)\times|\mathsf{x}|}$, $\mathsf{v}_j \in \mathbb{R}^{|\mathsf{x}|}$. By imposing further sparsity constraints on U_j and v_j so that $p(\mathsf{U}_j) \propto \exp\{-\gamma_U\|\mathsf{U}_j\|_1\}$ and $p(\mathsf{v}_j) \propto \exp\{-\gamma_v\|\mathsf{v}_j\|_1\}$, it is possible to identify features predictive of the underlying structures. As before, we assume that each expert is sparse in the augmented space of the visible and hidden variables $p(\mathsf{X}_j) \propto \exp\{-\gamma\|\mathsf{X}_j\|_1\}\mathbb{I}(\mathsf{X}_j \succ 0)$.

In this paper we consider $f(\mathsf{v}_j;\mathsf{x}^{(i)}) = \mathsf{v}_j^T\mathsf{x}^{(i)}$. Another useful parameterization is to let $f(\mathsf{v}_j;\mathsf{x}^{(i)}) = \mathsf{v}_j^T K_j(.,\mathsf{x}^{(i)})$ for a positive semi-definite kernel function K_j evaluated at n training points. By allowing $K_j(\mathsf{x}^{(i)},\mathsf{x}^{(j)}) \propto \exp\{-(\mathsf{x}^{(i)}-\mathsf{x}^{(j)})^T\mathsf{M}_j(\mathsf{x}^{(i)}-\mathsf{x}^{(j)})\}$ and imposing sparsity constraints on $\mathsf{v}_j \in \mathbb{R}^n$ and the diagonal matrix $\mathsf{M}_j \in \mathbb{R}^{|\mathsf{x}|}$, it may be possible to cluster structures based on subsets of features and data points.

For discriminative mixtures, objective (3) is redefined as

$$\mathcal{Q}_m = \langle \log p(\{\mathsf{y},\mathsf{z},\mathsf{w},\theta\}|\{\mathsf{x}\})\rangle_{p(\mathsf{w},\mathsf{z}|\{\mathsf{x}\},\{\mathsf{y}\};\theta_{old})}, \quad (8)$$

where θ_{old} defines the previously estimated parameters. Objective (8) needs to be optimized with respect to parameters θ that include the structures of experts X_j and parameters U_j, v_j (and M_j) for each mixture component j. The procedure for fitting discriminative mixtures of sparse latent Gaussians that we call *MSLICE* is a straightforward extension of Algorithm 1, where e.g.

$$\langle w_j^{(i)}|\mathsf{x}^{(i)},\mathsf{y}^{(i)}\rangle \propto \pi_j(\mathsf{x}^{(i)})\mathcal{N}_j(\mathsf{y}^{(i)};\mathsf{U}_j^y\mathsf{x}^{(i)},\mathsf{W}_{yy}^j), \quad (9)$$

and the rest is expressed analogously. Here $\mathsf{W}^j \stackrel{\text{def}}{=} [\mathsf{W}_{yy}^j\mathsf{W}_{yz}^j;\mathsf{W}_{zy}^j\mathsf{W}_{zz}^j] \stackrel{\text{def}}{=} \mathsf{X}_j^{-1}$ is the estimate of the covariance of the j^{th} expert, and $\mathsf{U}_j^{\mathsf{y}} \in \mathcal{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ is the y,x block of U_j. The updates for U_j and v_j are computed by a coordinate-wise gradient ascent on $\mathcal{Q}_m$. They are derived straightforwardly, in a similar way to the EM treatment of the classic LASSO regression [13].

Note that our conditional extension of the sparse latent field model vaguely resembles a CRF (e.g. [23]) by allowing for rich structures in the target variables and accommodating the conditioning on side information. However, in contrast to CRFs, it allows for missing and hidden targets, and imposes sparsity constraints on the underlying representations $p(\mathsf{y},\mathsf{z}|\mathsf{x})$. In this paper we focus on a simple setting where the couplings X_j of each expert depend on the external features x only through the choice of the mixing component. Richer sparse conditional latent fields may potentially be considered, e.g. by letting the edge-specific penalties γ_{ij} depend on x.

Other structured representations: By allowing for the link-specific penalties γ_{ij}, it is easy to encode prior knowledge about the structure. One limited special case is sparse factor analysis, where there is no residual structure in $p(\mathsf{y}|\mathsf{z})$, i.e. $\gamma_{y_iy_j} \to \infty$ for $i \neq j$. The results may be straightforwardly extended to more structured representations, such as trees or models with deep hierarchies (e.g. by setting $\gamma_{z_iz_j} \to \infty$ if z_i and z_j are in non-neighboring layers).

2.4 Non-Gaussian Marginals and Gaussianization (CopMSLICE)

SLICE trains a sparse Gaussian model, but stock price data is known to be non-Gaussian [12]. For example, in the FTSE100 data, the kurtoses of single stock distributions computed over the training time period varies in the range from 6.3 to 76.7 suggesting heaviness of the tails. Learning the structure of a general high-dimensional non-Gaussian distribution $p(\tilde{\mathsf{y}})$ is difficult. One way to do this is by decomposing the problem into two simpler tasks: capturing non-Gaussianity of the observations by separately learning low-dimensional non-Gaussian marginals, and modeling the high-dimensional dependency structure by a distribution from a tractable family.

A method of this type is to obtain "Gaussianized" [7] representations y_j of the non-Gaussian observations $\tilde{y}_j$, so that

$$y_j = \langle \tilde{y}_j \rangle + \sigma(\tilde{y}_j) F_G^{-1}(F_{\tau_j}(\tilde{y_j})), \qquad (10)$$

and fitting a Gaussian $\mathcal{N}(\mathsf{y})$ to the transformed data. Here F_G^{-1} is the inverse CDF of a standard Gaussian. $F_{\tau_j}(\tilde{y_j})$ is a monotonic approximation of the univariate marginal CDF of a non-Gaussian observation, estimated from data $\tilde{y}_j$ by fitting parameters τ_j. Learning the marginal CDFs $F_{\tau_j}(\tilde{y_j})$ is a relatively straightforward univariate modeling problem. Many potential methods are applicable. For example, [28] suggest using Winsorized approximations. Alternatively, we can choose to use kernel density estimation for its speed and simplicity in the "body" of the distribution, and separate generalized Pareto distributions [4] for each of the upper and lower tails.

The performance of GLASSO for the Gaussianized variables y is discussed in the very recent work of [28], who model the dependencies by the sparse precision $\mathrm{cov}^{-1}(\mathsf{y})$. The CDF of the resulting distribution over F_{τ_j}'s is a sparse Gaussian copula [28, 32]; we therefore refer to the method as CopGLASSO. We extend [28] to the latent variable setting and use the structure of $\mathrm{cov}^{-1}(\mathsf{y}, \mathsf{z})$ to model the dependencies. We use SLICE to train a multivariate Gaussian model of the transformed data. The resulting model combines the learned univariate CDFs with the structure defined by a sparse latent Gaussian field. To draw the analogy with CopGLASSO, we refer to the method as CopSLICE.

We extend MSLICE to CopMSLICE in a similar manner: we assign a Gaussianizing function to each expert in the mixture. The objective becomes

$$\mathcal{Q}_f = \mathcal{Q}_m + \sum_{i,k} \langle w_k^{(i)} | \mathsf{x}^{(i)}, \mathsf{y}^{(i)} \rangle \sum_j \log |y'_{jk}(\tilde{y}_j^{(i)})| \quad (11)$$

where $\mathcal{Q}_m$ is the MSLICE objective (8), i indexes data points, j indexes dimensions, k indexes experts, and y_{jk} is defined similarly to (10). Note that we need smooth and heavy-tail approximations of the experts' CDFs $F_{\tau_{jk}}$ to account for outliers and ensure differentiability in (11). We incorporate training of the marginals into MSLICE's EM algorithm by performing a gradient ascent for τ_{jk} within the M-step using the generalized Pareto approach.

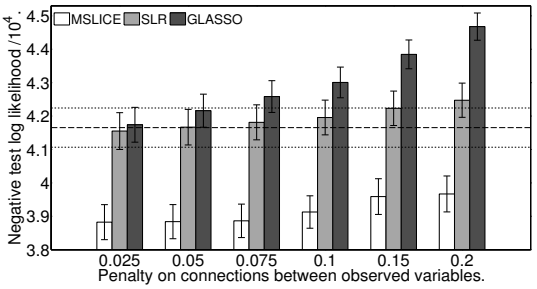

Figure 2: Negative test likelihoods of GLASSO, SLR, and MSLICE on the full set of FTSE100 composites, using ~ 1100 training and ~ 550 testing samples. The test error of the empirically estimated precision (with its error bars) is shown by the dashed line. MSLICE outperforms the competing models for all considered levels of sparsity.

3 Demonstration

We have applied (M)SLICE to financial stress-testing and visualization of dependencies between asset returns for the composites of the FTSE100 index, discarding stocks whose data did not cover the 6-year period. Our experimental settings are explanatory rather than predictive: we are trying to understand the past rather than predict the future. First we collected the daily closing prices of the stocks over the period from 2005 to 2011 from Yahoo Finance. The price data was cleaned and appropriately adjusted for splits, stock issues, and dividends. The data was converted to equity returns by computing the ratio of closing prices on consecutive days. There were 1633 days of returns data for each stock, which were sub-sampled at uniform random to generate lower-frequency samples. For illustration of the comparison of SLICE and GLASSO, it was assumed that temporal couplings between lower-frequency equity returns were negligible compared with the high-frequency intra-day fluctuations ($p(\mathsf{y}^{(t)}, \mathsf{z}^{(t)}, \mathsf{y}^{(t+\Delta)}, \mathsf{z}^{(t+\Delta)}) \approx p(\mathsf{y}^{(t)}, \mathsf{z}^{(t)}) p(\mathsf{y}^{(t+\Delta)}, \mathsf{z}^{(t+\Delta)})$ for large Δ). In the demonstration of MSLICE, it was assumed that the residual temporal couplings are explained by the technical indicators used as the side information, i.e. $p(\mathsf{y}^{(t)}, \mathsf{y}^{(t+\Delta)} | \mathsf{x}^{(t)}, \mathsf{x}^{(t+\Delta)}) = p(\mathsf{y}^{(t)} | \mathsf{x}^{(t)}) p(\mathsf{y}^{(t+\Delta)} | \mathsf{x}^{(t+\Delta)})$.

In our first experiment we compared GLASSO [14], the very recent sparse-low-rank (SLR) decomposition of [6] (see Section 4 for details), and MSLICE using all composites of FTSE100. All the competing methods need to set penalty parameters γ, and we computed test likelihoods for different such settings. Basically, we treated γ as a design

parameter, which is similar to [14, 28] and is a practical requirement of explanatory structure learning, where it is important to ensure that the model gives an accurate representation of data even if the model is very sparse. We used cross-validation to set the cardinality of the latent space $|\mathsf{z}|$ and the number of mixture components K for MSLICE, as both are low-dimensional discrete parameters. We looked at the range 1:5 for both K and $|\mathsf{z}|$ and found that values larger than 3 did not make much difference. We also tried using the SVD of the low-rank component in the SLR decomposition of $\mathbf{\Sigma}_{yy}^{-1}$ to set $|\mathsf{z}|$, but this made little difference. In the future, we will use nonparametric Bayesian methods to learn both; however, we are unsure whether this is going to make much difference for sparser models, as larger sparsity penalties will tend to prune extraneous components and dimensions. For MSLICE, a mixture of factor analyzers was used to initialize the gating parameters and means, while the initial precision of each expert was set to the precision learned by a single sparse Gaussian fitted to the training set. We only report the results for the non-kernelized parameterization of the gating distribution (6), with $f(\mathsf{v};\mathsf{x}) = \mathsf{v}^T\mathsf{x}$.

Figure 2 shows the negative test log-likelihoods of the models for 82 retained composites. We have used the first 1100 days for training, and the remaining 550 days for testing. As the side information x for MSLICE, we considered several volatility and trend indicators and indices computed over the preceding training samples for a different market (SnP500). For SLR and GLASSO, we modeled the side information x jointly with the FTSE100 composites y, and used the conditional $p(\mathsf{y}|\mathsf{x})$ in the test likelihood computations. From Figure 2 we see that for all the considered sparsity parameters, MSLICE significantly outperforms GLASSO and SLR *independently* of the degree of sparsity. Interestingly, the single most relevant feature important for predicting the structural changes in MSLICE was the implied volatility index (VIX).

We then investigated the non-Gaussian extensions of the models CopGLASSO, CopSLICE, and CopMSLICE. For this experiment, we used the 20 largest capitalization stocks from the financial, mining, and consumable sectors. As before, K and $|\mathsf{z}|$ were determined by cross-validation. In one experiment, in order for the comparison to be favorable to GLASSO and SLR, the data was randomly sub-sampled from the 400-day low-volatility period that was found to be well-modeled by a single mixture component. The models were also evaluated on low-volatility test data. For the demo shown in Figure 3 (top), for each limit on the number of non-zero links, we started with a low γ and iteratively increased it until each model satisfied this sparsity constraint. We see that CopSLICE is the best performing model across a broad range of sparsity constraints. The copula extension of GLASSO that does not use latent variables performs well for dense models, but breaks when the

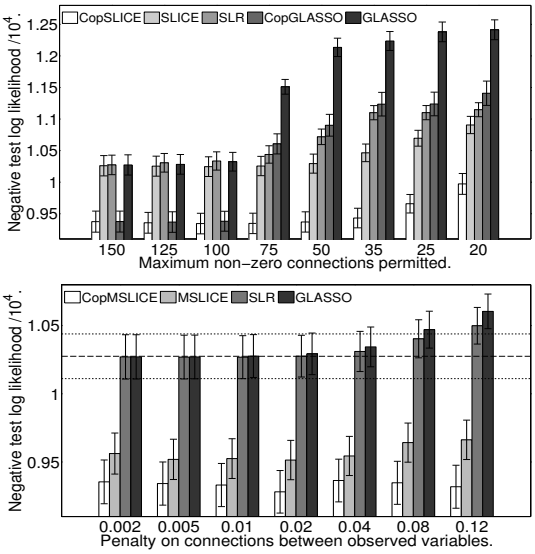

Figure 3: *Top:* test errors of the sparse copula and other unimodal models for low-volatility periods. The number of possible links is 231. *Bottom:* test errors of the multimodal models, GLASSO, and SLR for combined low- and high- volatility periods. The test error of the empirically estimated precision (with error bars) is shown by the dashed line. CopMSLICE, MSLICE, and CopSLICE are the best performing models according to the criteria.

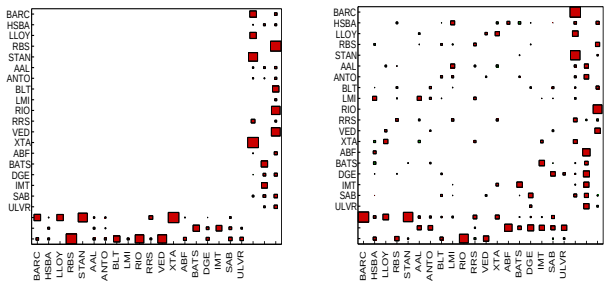

Figure 4: Visualization of relationships between composites of FTSE100 using 2-component MSLICE corresponding to the low- and high-volatility periods (left, right). The last three columns represent latent factors. Note the denser structure of the higher-volatility component.

number of links becomes small. We then repeated the previous experiment by using the full range of low- and high-volatility data (see Figure 3 (bottom)), computing the test errors on the out-of-sample observations. We can again see that our conditional mixture approaches, CopMSLICE and MSLICE, are the best performers across all the considered sparsity parameters.

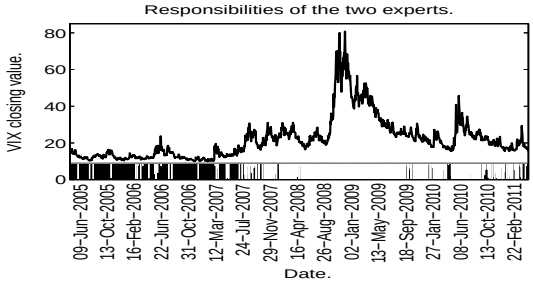

Figure 5: The VIX (top) plotted against the posterior responsibility $p(w^{(t)}|\mathsf{x}^{(t)}, \mathsf{y}^{(t)})$ for the 2-component mixture (bottom). The VIX was the single most important feature for predicting changes in structure.

Structure Interpretation

After fitting discriminative mixtures of sparse latent Gaussian fields, we proceeded to interpret the resulting structures $p_w(\mathsf{y}, \mathsf{z})$ and mixing variables w. We have used $K = 2$, $|\mathsf{z}| = 3$, and set γ by cross-validation for this visualization experiment. Figure 4 shows precisions of sparse latent Gaussians $p_w(\mathsf{y}, \mathsf{z})$ corresponding roughly to the low- and high-volatility periods. The latent factors z roughly capture sector information (the financials, mining, and consumables are sorted from left to right). RBS's link to the commodities could be explained by its over-exposure to the coal-mining industry. Note that there are many residual dependencies between the banks and the mining companies in the high-volatility expert. This is to be expected during the recession period after the financial crisis began to affect the economy (thus lowering the demand for materials). The "defensive" consumables are largely unaffected, with the two strongest pairs in $p_w(\mathsf{y}|\mathsf{z})$ being the two drinks (SAB, DGE) and tobacco (BATS, IMT) companies. This illustrates the spectacular collapse of the so-called "decoupling theory" in 2008, which stated that developed markets (largely built around financial services) are "decoupled" from the emerging commodity-oriented markets. Figure 5 shows the posterior responsibility $p(w^{(t)}|\mathsf{x}^{(t)}, \mathsf{y}^{(t)})$ for the 2-component mixture (bottom) plotted against the most important side feature – the temporally delayed VIX, where the black and the white bars at the bottom indicate the probability of being in either one or the other state.

Financial Stress-Testing

The idea behind financial stress-testing is to test robustness of a fund or company under fantasy worst-case scenarios, such as an extended period of extraordinarily high volatility, collapse or jump of commodity prices, unavailability of credit, exceptionally high pressure on a sector index and likewise, as well as combinations of multiple shocks. In order to test positioning of a fund's portfolio, it is important to be able to generate fantasy equity returns that may be reflective of the modeled scenario. After the crisis of the late 2000s, stress-testing is playing an increasingly important role in financial due diligence and is becoming a regulatory requirement in Europe. The common approach to stress-testing relies heavily on human expertise, heuristically assigning higher weights to historical patterns characterized by given combinations of shock factors ([1, 26]). Such approaches may be prone to human error and are difficult to analyze or extend. Moreover, it may be difficult to identify subsets of observations corresponding to combinations of generally continuous factors, as many of the extreme conditions may have been observed rarely in the past. A natural alternative approach to generating fantasy data corresponding to a shock scenario is by using our MSLICE approach, which may be applicable even if the number of past observations is small compared with the problem's dimensionality. By including external shocks as side information in $p(\mathsf{y}|\mathsf{x})$, it may be possible to generate numerous structures for combinations of fantasy scenarios.

In our simple stress-testing experiment, we generated 1000 points of fantasy VIX data by running a Brownian motion simulation, centering the output on the high value of 60, and then clipping it to the range 40-80. Samples from the two trained models (MSLICE and GLASSO) were generated conditioned on this fantasy VIX data; for GLASSO, we again modeled the joint $p(\mathsf{y}, x)$ and sampled fantasy data from $p(\mathsf{y}|x)$ where x was VIX. The other model was a Gaussian fitted to a manually labeled subset of training points from a high volatility region. As a proxy for the ground truth, we used another Gaussian $p_{GT}(\mathsf{y})$ with the empirical mean and covariance computed at the out-of-sample high volatility regions. For each model we estimated $KL(p_{GT}(\mathsf{y})\|p(\mathsf{y}|x \in \mathcal{R}))$, where $\mathcal{R}$ is the space of generated shock scenarios. We have used cross-validation to set γ, K, and $|z|$ for MSLICE, and γ for GLASSO. We found the divergences of $KL_{mslice} \approx \mathbf{4.35 \pm 0.77}$, $KL_{glasso} \approx 5.97 \pm 1.44$, and $KL_{man} \approx 4.77 \pm 0.89$ for MSLICE, GLASSO, and the manually generated models, with the variance due to multiple subsamples. Neither MSLICE nor GLASSO required human expertise, and used the labelings only for validation. While validations of stress-testing results are complex and market-dependent, it is encouraging that the model was able to outperform a human expert. This suggests that MSLICE may be a promising tool for generating fantasy scenarios for stress testing.

4 Discussion

We have described a simple and efficient approach for learning structures of sparse latent field models based on the assumption of sparsity in the augmented space of visible and hidden variables $\{\mathsf{y}, \mathsf{z}\}$. Our approach replaces a combinatorial search over model structures by a continuous optimization in the space of all such structures, with sparsity constraints on the graphical models corresponding

to the joint distributions $p(\mathsf{y}, \mathsf{z})$. It may be useful for a variety of high-dimensional real-world scenarios, where the present incomplete patterns $\{\mathsf{y}^{(t)}, \mathsf{z}^{(t)}\}$ may be affected by the past data $\mathsf{x}^{(t-1)}$, but little is known *a priori* about the structural family of the underlying model.

We extended the recently introduced methods for learning sparse structures of fully observed Gaussians [5, 14, 10, 34, 24] to handle latent variables and side information. There has been recent work on other methods for inferring structures of Gaussian graphical models. For example [2] describes a generative model for structures of fully observed Gaussian models, where each observed variable y_i belongs to a latent class $z_i \in \{1, \ldots, k\}$, and the edge between y_i and y_j is penalized by a term $\gamma_{z_i z_j}$. They describe a variational approximation of the inference and comment on the technical difficulties of setting these penalty parameters. [3] formulates a more general approach for fully observed pairwise Markov networks, with binomial priors on the present or absent links. Our framework is different, since we are interested in latent variable models with sparsity in $\{\mathsf{y}, \mathsf{z}\}$, rather than fully visible networks with sparsity in $\{\mathsf{y}\}$. For the Gaussians, we avoid the need to resort to variational approximations of the posteriors of class memberships or present/absent links, which makes the method scalable to high-dimensional problems with latent variables.

The approach most closely related to ours is the *sparse/low-rank* (SLR) decomposition of [6]. Their method relies on the assumption of sparsity of the Gaussian conditional $p(\mathsf{y}|\mathsf{z})$, where y and z correspond to the data and latent variables respectively. They note that if the dimensionality of the latent factors $|\mathsf{z}|$ is low, the marginal precision of the observed variables decomposes into sparse and low-rank positive semi-definite matrices. They proceed by defining a convex optimization problem on the regularized likelihood for the sparse and low-rank components. Because the approach does not explicitly parameterize $p(\mathsf{y}, \mathsf{z})$, it requires solving a separate sparse matrix factorization problem if a latent-space visualization is desired. The SLR approach is elegant and computationally attractive; however, in contrast to our formulation, SLR is an optimization algorithm rather than a probabilistic graphical model, where encoding of the prior knowledge, extensions to mixtures, or inclusion of the side information is not straightforward. Another recent extension of GLASSO [37] allows for missing observations without systematically hidden variables. In contrast to [37], our approach is applicable for both missing observations and hidden variables, and is extended to discriminative non-Gaussian mixtures. We have demonstrated that such extensions help to significantly improve on SLR [6] and other Gaussian [5, 14] or Gaussian copula [28] methods in terms of test likelihoods.

The presence of the discrete latent states makes the optimization surface of MSLICE and its copula version non-convex. However, it may be efficiently optimized via a simple iterative procedure which we demonstrated empirically to have excellent, robust modeling performance on test data. Rather than formally analyzing the performance of convex optimization methods for sub-optimal sparse Gaussian models, our goal here was to introduce and empirically demonstrate the effectiveness of an intrinsically non-convex approach that may be more effective for uncovering structure in real-world financial data.

Our generic formulation of MSLICE is very simple, and many interesting extensions touched upon only briefly will be explored further in the future. At present, temporal information has only been included by conditioning on temporally dependent features, such as technical indicators of longer-term trends or market volatility. Our future work will consider extensions of the model to non-Gaussian latent field models with sparse structures in the augmented space $\{\mathsf{y}, \mathsf{z}\}$ that may be more suitable for modeling dependencies between equity returns in financial applications. The model will be extended to explicitly capture temporal and spatial dependencies between the latent factors and observations, with sparsity constraints on transitions.

In this work, we selected the number of mixture components by cross-validation. In the future we will develop infinite mixtures [33] of sparse latent field experts, and consider other Bayesian nonparametric methods [17, 22] for learning the number of latent factors. We will also investigate priors favoring group sparsity, and consider more efficient optimization approaches in the M-step. Applications will focus on other areas of financial risk management including portfolio construction, as well as other hot areas of network modeling including biomarker discovery and patient stratification in personalized medicine.

Finally, we note that (M)SLICE combines the higher-level problem of discovering an underlying structural form with the lower-level problem of identifying the specific instance of that structure that produces the best explanation of the observations. Similarly to GLASSO, our approach is applicable for a large-p small-n setting; it is also tractable enough to be used for moderately high-dimensional datasets with hidden variables and/or missing observations. It may be effective for financial stress-testing and knowledge discovery. We showed that when applied to financial stress-testing, our method does not need to rely on expensive human expertise. We also discussed an application to the study of relationships between equity returns under varying market regimes. More generally, the approach may have important applications to systems biology, social marketing, and financial risk minimization.

Acknowledgements

Felix Agakov acknowledges MRC (G0800604) for its support in the Centre for Population Health Sciences, University of Edinburgh.

References

[1] C. Alexander. *Market Models: A Guide to Financial Data Analysis*. Wiley, 2001.

[2] C. Ambroise, J. Chiquet, and C. Matias. Inferring sparse Gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3, 2009.

[3] S. Amizadeh and M. Hauskrecht. Latent variable models for learning in pairwise Markov networks. In *AAAI*, 2010.

[4] A. Balkema and L. de Haan. Residual life time at great age. *Annals of Probability*, 2, 1974.

[5] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9, 2008.

[6] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. Unpublished, 2010.

[7] S. Chen and R. A. Gopinath. Gaussianization. In *NIPS*, 2000.

[8] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14, 1968.

[9] A.P. Dempster. Covariance selection. *Biometrics, Special Multivariate Issue*, 28, 1972.

[10] J. Duchi, S. Gould, and D. Koller. Projected subgradient methods for learning sparse gaussians. In *UAI*, 2008.

[11] G. Elidan, I. Nachman, and N. Friedman. "Ideal parent" structure learning for continuous variable Bayesian networks. *Journal of Machine Learning Research*, 8:1799–1833, 2007.

[12] E.F. Fama. The behavior of stock-market prices. *Journal of Business*, 38(1), 1965.

[13] M. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Trans. on PAMI*, 25(9), 2003.

[14] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 2008.

[15] J. Friedman, T. Hastie, and R. Tibshirani. Applications of the Lasso and grouped Lasso to the estimation of sparse graphical models. Technical report, Stanford University, 2010.

[16] N. Friedman, I. Nachman, and D. Peer. Learning Bayesian network structure from massive datasets: the "Sparse Candidate" algorithm. In *UAI*, 1999.

[17] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *NIPS*, 2005.

[18] S. Harmeling and C. Williams. Greedy learning of binary latent trees. *IEEE Trans. Pattern Mach. Intell.*, 33(6), 2011.

[19] D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[20] M. Jordan and R. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:415–447, 1992.

[21] C. Kemp and J. B. Tenenbaum. The discovery of structural form. *Proceedings of National Academy of Sciences*, 105(31), 2008.

[22] D. Knowles and Z. Ghahramani. Nonparametric Bayesian sparse factor models with application to gene expression modelling. *Annals of Applied Statistics*, 5, 2011.

[23] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[24] B. Lake and J. Tenenbaum. Discovering Structure by Learning Sparse Graphs. In *CogSci*, 2010.

[25] S. L. Lauritzen. *Graphical Models*. Oxford Uni Press, 1996.

[26] H. Leinonen, editor. *Simulation analyses and stress testing of payment networks, BoF E:42*. Bank of Finland, 2009.

[27] E. Levina, A. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, 2(1):245–263, 2008.

[28] H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *JMLR*, 10, 2009.

[29] M. H. Maathius, M. Kalisch, and P. Buhlmann. Estimating high-dimensional intervention effects from observation data. *The Ann. of Stat.*, 37:3133–3164, 2009.

[30] B. M. Marlin and K. P. Murphy. Sparse Gaussian Graphical Models with Unknown Block Structure. In *International Conference on Machine Learning*, 2009.

[31] N. Meinshausen and P. Buehlmann. High-dimensional graphs and variable selection with lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

[32] R. Nelsen. *An introduction to copulas*. Springer, second edition, 2006.

[33] C. Rasmussen and Z. Ghahramani. Infinite mixtures of Gaussian process experts. In *NIPS*, 2001.

[34] K. Scheinberg, S. Ma, and D. Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In *NIPS*, 2010.

[35] M. Schmidt, G. Fung, and R. Rosales. Fast optimization methods for L1 regularization: a comparative study and two new approaches. In *Machine Learning: ECML 2007*, volume 4701 of *LNCS*. Springer, 2007.

[36] R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7, 2006.

[37] N. Staedler and P. Buehlmann. Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, 2010. DOI 10.1007/s11222-010-9219-7.

[38] K.C. Toh, M. J. Todd, and R.H. Tütüncü. SDPT3 – a Matlab software package for semidefinite programming. *Optimization Methods and Software*, 11:545–581, 1999.

[39] C. Wang, D. Sun, and K. C. Toh. Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm. *SIAM Journal on Optimization*, 20, 2010.

[40] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1), 2007.

[41] N. L. Zhang. Hierachical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5, 2004.