Large-scale Probabilistic Forecasting in Energy Systems using Sparse Gaussian Conditional Random Fields

Matt Wytock and J. Zico Kolter

***Abstract*— Short-term forecasting is a ubiquitous practice in a wide range of energy systems, including forecasting demand, renewable generation, and electricity pricing. Although it is known that probabilistic forecasts (which give a distribution over possible future outcomes) can improve planning and control, many forecasting systems in practice are just used as "point forecast" tools, as it is challenging to represent high-dimensional non-Gaussian distributions over multiple spatial and temporal points. In this paper, we apply a recently-proposed algorithm for modeling high-dimensional conditional Gaussian distributions to forecasting wind power and extend it to the non-Gaussian case using the copula transform. On a wind power forecasting task, we show that this probabilistic model greatly outperforms other methods on the task of accurately modeling potential distributions of power (as would be necessary in a stochastic dispatch problem, for example).**

I. Introduction

Forecasting, the task of predicting future time series from past observations, is ubiquitous in energy systems. As well-known examples, electricity system operators routinely forecast upcoming electrical load and use these forecasts in market planning [17], [19]; wind farms forecast future power production when offering bids into these markets [7], [11], [15]; and there is a growing use of forecasting at the micro-scale for coordinating smart grid operations [1]. Despite their ubiquity and the complexity of many forecasting methods, most methods are ultimately employed as "point forecast" strategies; users train a system to output point predictions of upcoming values, typically to minimize a metric such as root mean squared error. However, for many complex control and planning tasks, such point forecasts are severely limited: the processes that make up electrical demand, wind power, etc, are stochastic systems and the notion of a "perfect forecast" is unattainable. Thus, *probabilistic forecasts*, which output a distribution over potential future outcomes instead of a single prediction, are of substantial practical interest. Indeed, studies have demonstrated that in the context of electrical demand and wind power, probabilistic forecasts can offer substantial benefits over point predictions [14].

Unfortunately, the challenge of probabilistic forecasts is that it is often very hard to describe the joint distribution over all predicted values because many variables of interest are highly non-Gaussian and it can be difficult to accurately model correlations in a high-dimensional output space. For this reason, most of the literature on probabilistic forecasting has often made simplifying assumptions, for example using specific forms of Gaussian linear models, such as autoregressive moving average (ARMA) models (e.g. [10]); or only predicting non-Gaussian marginal distributions for single output variables (e.g. [9]). Indeed, past work has explicitly highlighted the challenge of developing models that can capture joint distributions over future values.

In this paper, we apply and extend a recently-proposed method for learning high-dimensional conditional Gaussian distributions, called sparse Gaussian conditional random fields, resulting in a method that is able to both capture high-dimensional correlations between a very large number of output variables and model non-Gaussian marginal probabilities. We accomplish this by 1) applying recent techniques in machine learning and statistics, which model high-dimensional correlations by exploiting sparsity in the *inverse covariance matrix*, and by 2) using copula transforms to capture non-Gaussian marginal distributions. We apply this model to the task of wind power forecasting (a setting where capturing non-Gaussian distribution and temporal and spatial correlation is critical), and show that it is able to make probabilistic predictions far better than forecasts that do not explicitly model spatial and temporal correlation. We also highlight the applicability of the method to challenging open problems in this area such as wind power ramp prediction.

II. The probabilistic forecasting setting

We consider the following setting: let $z_t \in \mathbb{R}^n$ denote a *vector-valued observation* at time t; for example, the ith element of z_t, denoted $(z_t)_i$ could denote the power output by a particular wind farm at time t, and i could range over a collection of wind farms. We let $w_t \in \mathbb{R}^m$ denote a set of (known) exogenous variables that may affect the evolution of the sequence; for example, w_t may include the current time of day, day of the year, and even external variables such as wind forecasts at upcoming time points. The goal of our forecasting setting is to predict H_f future values given H_p past values and the exogenous variables:

$$\text{Given } z_{t-H_p+1:t}, w_t, \text{ predict } z_{t+1:t+H_f}. \tag{1}$$

This setting can be referred to as the *vector autoregressive exogenous* (VARX) setting [13]; however, this terminology is also used to describe a particular form of probabilistic model for the sequence, so we just refer to it generally as a multivariate forecasting problem.

Although predicting a single estimate of future observations from past observations can be very useful in many situations, we often want to understand more broadly the

M. Wytock and J. Z. Kolter are with the School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA {mwytock,zkolter}@cs.cmu.edu

distribution of possible future observations given past observations and exogenous factors, denoted as

$$p(z_{t+1:t+H_f}|z_{t+1:t+H_f}, w_t). \tag{2}$$

We are particularly interested in the case where individual observations z_t are high-dimensional and we want to predict their evolution over a relatively long time horizon, resulting in a high-dimensional probability distribution. The task is made more challenging by the fact that the observations may not have Gaussian distributions (indeed, in the case of our wind power setting, they typically do not) and by the fact that we may have relatively few past observations upon which to build our high-dimensional model.

A. Relation to existing settings and models

A common method for handling such settings is a vector autoregressive exogenous model in which we model future observations as a linear combination of past observations, exogenous variables, and a Gaussian noise term

$$z_{t+1} = \sum_{i=0}^{H_p-1} \Theta_i z_{t-i} + \Psi w_t + \epsilon_t \tag{3}$$

where $\Theta_i \in \mathbb{R}^{n\times n}$ and $\Psi \in \mathbb{R}^{n\times m}$ are model parameters, and $\epsilon_t \sim \mathcal{N}(0, \Sigma)$ is a zero-mean Gaussian random variable with covariance Σ. If we want to make forecasts over a multiple future time points, we can iteratively apply this model or explicitly build an additional VARX model to predict z_{t+2} from past values.

A common extension to the autoregressive framework is to add a moving sum of noise variables, resulting in the autoregressive moving average (ARMA) model (these and similar extensions are described in several sources, e.g. [3]). Additionally, we may choose Θ such that the overall system has at least one unit root (the ARIMA model) or consider only a certain periodic set of past observations (the seasonal ARIMA model). These and other approaches have been used extensively in the literature, and while they do impose a joint probabilistic model over future observations, they are very limited in the form of this distribution (multivariate Gaussian with a particular covariance matrix).

In the sequel we will consider forecasting using a model that allows for more general dependencies between predicted variables, can capture non-Gaussian marginal distribution of the variables via a copula transform, and which can be learned from very little data by exploiting sparsity. We focus largely on extensions of the pure autoregressive setting, but the model can also be extended to the ARMA setting by introducing additional latent variables.

III. Forecasting with the sparse Gaussian conditional random field

Here we describe the sparse Gaussian conditional random field (SGCRF), a model that has has recently been proposed by several authors [16], [21], including our own work [20] which develops the algorithm we apply here. In this paper, we focus on the extension of this method to the case of non-Gaussian output variables and the application to probabilistic forecasting in energy systems. For a full description of the algorithm and theoretical analysis, see [20] which focuses on the pure machine learning aspects of the model.

For simplicity of notation, we refer to the set of all known variables as a single vector $x \in \mathbb{R}^n$, while the unknown variables we are attempting to predict are given by $y \in \mathbb{R}^p$

$$x = \begin{bmatrix} z_t \\ z_{t-1} \\ \vdots \\ z_{t-H_p+1} \\ w_t \end{bmatrix}, \quad y = \begin{bmatrix} z_{t+1} \\ z_{t+2} \\ \vdots \\ z_{t+H_f} \end{bmatrix}. \tag{4}$$

In this section we first present the pure SGCRF formulation, which models $y|x$ as a Gaussian distribution, exploiting sparsity to capture high-dimensional correlation using very few model parameters. We then discuss the extension to non-Gaussian distributions via the copula transform and summarize the algorithm.

A. The pure SGCRF model

The SGCRF models $y|x$ as a multivariate Gaussian distribution in terms of its exponential family form, via *precision* (inverse covariance) and *mean parameters* rather than a covariance and mean. The advantage of this formulation is *sparsity*: in many forecasting domains, virtually all variables share some degree of correlation (for instance, the wind ten hours from now may be correlated with the wind one hour from now, in a purely statistical sense) and thus modeling the full joint distribution requires a large number of parameters. However, it is well-known that the *inverse* of the covariance matrix for a Gaussian captures just the *conditional independencies* of the distribution and can be highly sparse even when the covariance matrix is dense. In the non-conditional case, this fact has been exploited by several authors to formulate efficient convex methods for sparse inverse covariance estimation using the ℓ_1 penalty [2], also known as the graphical lasso [5].

For prediction tasks such as probabilistic forecasting, it has repeatedly been observed that a discriminative approach modeling the conditional distribution $(y|x)$ can be superior to the generative approach of modeling the full distribution (y, x) [12]. Thus, we extend sparse inverse covariance estimation to the discriminative case, jointly modeling the correlations in the output variables along with their conditional dependence on input features. Since the discriminative setting coincides with the standard notion of a conditional random field [18], we refer to this model as the sparse Gaussian conditional random field (SGCRF).

Formally, the SGCRF models the distribution $y|x$ as

$$p(y|x; \Theta, \Lambda) = \frac{1}{Z(x)} \exp\left\{-\frac{1}{2} y^T \Lambda y - x^T \Theta y\right\} \tag{5}$$

where $\Lambda \in \mathbb{R}^{p\times p}$ and $\Theta \in \mathbb{R}^{n\times p}$ are the parameters of the model, and $Z(x)$ is the partition function (used to ensure the distribution over y integrates to one) given by

$$\frac{1}{Z(x)} = c|\Lambda|^{1/2} \exp\left\{-\frac{1}{2} x^T \Theta \Lambda^{-1} \Theta x\right\} \tag{6}$$

where c is a constant term that is independent of x, Λ, and Θ. Critically, we can express models with high correlation between variables even though Λ and Θ are sparse. To see this, note that the model can easily be transformed to mean/covariance form

$$p(y|x) \sim \mathcal{N}(-\Lambda^{-1}\Theta^T x, \Lambda^{-1}) \tag{7}$$

but $\Lambda^{-1}\Theta$ and Λ^{-1} are likely dense even when Λ and Θ are sparse. Thus, in the forecasting setting, each element of our prediction $z_{t+1:t+H_f}$ can depend on every element of $z_{t-H_p+1:t}$ and w_t. By exploiting sparsity, we learn the model efficiently from much less data than would be required to estimate the mean and covariance directly.

B. A second-order optimization method for the SGCRF

Next, we describe an efficient second-order optimization method for the SGCRF which exploits the sparse structure of the solution with an active set approach. In particular, given a collection of input-output pairs $\{x_i, y_i\}$, $i = 1, \ldots, m$, we estimate the parameters Λ and Θ using maximum likelihood and encourage sparsity by adding an additional ℓ_1 penalty $\lambda(\|\Lambda\|_1 + \|\Theta\|_1)$, where λ is a regularization parameter and $\|\cdot\|_1$ denotes the elementwise ℓ_1 norm of a matrix. This results in the optimization problem

$$\begin{aligned} \underset{\Lambda,\Theta}{\text{minimize}} \quad & \log|\Lambda| + \operatorname{tr}\Lambda S_{yy} + 2\operatorname{tr}\Theta S_{yx} + \\ & \operatorname{tr}\Lambda^{-1}\Theta^T S_{xx}\Theta + \lambda(\|\Lambda\|_1 + \|\Theta\|_1) \end{aligned} \tag{8}$$

where S_{yy}, S_{yx} and S_{xx} are sample covariance matrices

$$S_{yy} = \frac{1}{m}\sum_{i=1}^m y_i y_i^T,\ S_{yx} = \frac{1}{m}\sum_{i=1}^m y_i x_i^T,\ S_{xx} = \frac{1}{m}\sum_{i=1}^m x_i x_i^T. \tag{9}$$

Although this is a convex problem, existing solvers suffer from slow convergence due in part to the coupling of the parameters in the matrix-fractional term $\operatorname{tr}\Lambda^{-1}\Theta^T S_{xx}\Theta$; in practice, forecasting over long time horizons across multiple locations quickly becomes intractable. Therefore, we develop a custom second-order method which is several orders of magnitude faster than standard solvers, enabling us to apply this method to large-scale probabilistic forecasting tasks such as forecasting wind power production over several adjacent wind farms and multiple days. Here we give the motivation and highlight the key technical points of the algorithm, but for complete details and theoretical analysis see [20].

The main idea is to decompose the objective function into a smooth term plus regularization term and find the descent direction by iteratively optimizing a regularized version of the second-order approximation to the smooth term. In general, for an optimization problem with the objective $f(x) + \lambda\|x\|_1$, the second-order Taylor expansion of f is

$$f(x+\Delta) \approx g(\Delta) \equiv f(x) + \nabla_x f(x)^T\Delta + \frac{1}{2}\Delta^T\nabla_x^2 f(x)\Delta \tag{10}$$

where $\nabla_x f(x)$ and $\nabla_x^2 f(x)$ denote the gradient and Hessian respectively. For a fixed x, finding Δ that minimizes this second-order expansion along with the ℓ_1 penalty is given by the solution to the regularized quadratic program

$$d = \arg\min_{\Delta} g(\Delta) + \lambda\|x+\Delta\|_1 \tag{11}$$

which we solve using coordinate descent. Given this direction, we choose our step size using backtracking line search and iterate until convergence. Note that our model is parameterized by two matrices Λ and Θ and thus the descent direction is pair of matrices, Δ_Λ and Δ_Θ.

Using coordinate descent for computing the descent direction is particularly appealing since it allows us to exploit sparsity in the solution by maintaining a small active set. In particular, even though Λ and Θ are comprised of $p(p+1)/2 + np$ different variables, due to the sparsity induced by the ℓ_1 penalty and the nature of forecasting applications, we expect the majority of these coordinates to be zero. We exploit this fact by considering only the set of variables that violate the optimality conditions; optimizing over Λ_{ij} (respectively Θ_{ij}) only if

$$\begin{aligned} |(\nabla_\Lambda f(\Lambda,\Theta))_{i,j}| > \lambda \text{ or } \Lambda_{ij} \neq 0 \\ |(\nabla_\Theta f(\Lambda,\Theta))_{i,j}| > \lambda \text{ or } \Theta_{ij} \neq 0. \end{aligned} \tag{12}$$

It can be shown that this heuristic produces an algorithm that converges to the optimal solution even though the descent direction at each step is not necessarily optimal. In practice, for problems with a sparse solution, this results in a substantial speed increase over the naive approach of considering every variable.

Finally, in order to make coordinate descent efficient, it is critical to cache and iteratively update certain matrix products. The gradient and Hessian of our objective are quite involved due to the $\operatorname{tr}\Lambda^{-1}\Theta^T S_{xx}\Theta$ term, however performing the coordinatewise updates efficiently reduces to maintaining the product of $\Delta_\Lambda\Lambda^{-1}$ and $\Delta_\Theta\Lambda^{-1}$. At each step in the coordinate descent inner loop, we reuse these products and when we update a single coordinate, $(\Delta_\Lambda)_{ij}$ ($(\Delta_\Theta)_{ij}$), we must update the corresponding ith row of $\Delta_\Lambda\Lambda^{-1}$ (respectively $\Delta_\Lambda\Lambda^{-1}$). It can be shown that the resulting algorithm achieves superlinear convergence, which in practice allows this model to be applied to many previously intractable probabilistic forecasting problems.

C. Non-Gaussian distributions via copula transforms

The above SGCRF is limited in that it can only model the distribution over y as a multivariate Gaussian. To overcome this limitation, we employ a (Gaussian) copula transform [22], a method for converting multivariate Gaussian distributions into multivariate distributions with arbitrary marginal distributions. Previous work has applied the copula transform to extend the sparse Gaussian MRF to non-Gaussian distributions [8] and here we extend this to the SGCRF, forming a model which is well-suited for probabilistic forecasting in a wide variety of energy systems.

Formally, suppose $u \in \mathbb{R}$ is a univariate random variable with cumulative distribution function (CDF) F; when we

only have samples of u, we use the empirical CDF

$$\hat{F}(u) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{u < u_i\}. \tag{13}$$

In the case that we expect the variables to come from known distribution (e.g. the Weibull distribution for modeling wind speeds), we could use the analytical CDF of this distribution directly. The copula transform simply converts the sample distribution to a uniform $[0, 1]$ random variable by the CDF F, then applies the inverse normal CDF Φ^{-1} to transform the $[0, 1]$ random variable into a Gaussian random variable. Our algorithm models the variables using a SGCRF in this transformed Gaussian space, and then transforms back to the original distribution by applying the inverse copula transform (the normal CDF Φ followed by the inverse CDF F^{-1}).

D. Final Algorithm

As with all learning methods, using SGCRFs consists of a training stage where we learn the parameters that maximize the model's likelihood on past observations. Then, for a new scenario (denoted $x' \in \mathbb{R}^n$), we use the model to make predictions about the future observations y'. The training stage consists of the following elements:

1) Given data (x_i, y_i), for $i = 1, \ldots, m$ (recall that each x_i consists of H_p past observations and external inputs w_t, and each y_i consists of H_f future observations), first estimate the univariate marginal distributions of each $(y_i)_j$, denoted F_j.
2) Transform each the y_i variables to a variable with marginal Gaussian distributions $\tilde{y}_i$ by applying the elementwise copula transform

$$(\tilde{y}_i)_j = \Phi^{-1}(F_j((y_i)_j)) \tag{14}$$

3) Train a SGCRF model (i.e., estimate the Θ and Λ parameters) on $(x_i, \hat{y}_i)$, $i = 1, \ldots, m$.

With a model, we can perform any of the following tasks:

- **Compute the most likely output:** Compute the mean in the Gaussian space $\tilde{y}' = -\Lambda^{-1}\Theta^T x'$; then transform each element of $\tilde{y}'$ using the inverse copula transform

$$(\hat{y}')_j = F^{-1}(\Phi((\tilde{y}')_j)) \tag{15}$$

- **Compute the probability of a given output y':** Convert y' to the Gaussian space using (14) and compute

$$\begin{aligned} p(y'|x') =& p(\tilde{y}'|x'; \Theta, \Lambda) \\ =& \frac{1}{Z(x')} \exp\left\{-\frac{1}{2}(\tilde{y}')^T\Lambda\tilde{y}' - (x')^T\Theta\tilde{y}'\right\} \end{aligned} \tag{16}$$

- **Draw a random sample of future observations:** Sample

$$\tilde{y}' \sim \mathcal{N}(-\Lambda^{-1}\Theta^T x', \Lambda^{-1}) \tag{17}$$

and then apply the inverse copula transform (15) to $\tilde{y}'$.

TABLE I
COMPARISON OF PREDICTION ERROR

Algorithm	RMSE
Linear Regression	0.1560
Linear Regression + copula	0.1636
ARMAX	0.1714
SGCRF	**0.1488**
SGCRF + copula	0.1584

IV. Application to wind forecasting

In this section, we describe the primary applied result of this paper, an application of the above probabilistic forecasting method to a real-world wind power prediction task. We use data from the GEFCom 2012 forecasting challenge, a wind power forecasting competition that was recently held on Kaggle [6], where the goal was to predict power output from 7 nearby wind farms over the next 48 hours using forecasted wind speed and direction as input variables.

In our setup, we model wind power production jointly across all wind farms as $z_t \in \mathbb{R}^7$ and include the forecasted wind at each farm as exogenous variables. We model the non-linear dependence of the wind power using radial basis functions (RBFs) with centers and variances tuned using cross-validation, resulting in 10 RBFs for each time point and location and $w_t \in \mathbb{R}^{3360}$. We also include autoregressive features for past wind power over the previous 8 hours which we found experimentally to be sufficient to capture the autoregressive behavior of wind power in this dataset. In our framework, the input and output variables (x_t, y_t) are compromised of w_t and z_t ranging over past and future time points

$$x_t = \begin{bmatrix} z_t \\ \vdots \\ z_{t-7} \\ w_t \end{bmatrix}, \quad y_t = \begin{bmatrix} z_{t+1} \\ \vdots \\ z_{t+48} \end{bmatrix} \tag{18}$$

resulting in $x_t \in \mathbb{R}^{3416}$ and $y_t \in \mathbb{R}^{336}$.

We fit the model using 80% of the provided data (874 training examples) and report results on the held out set. As baselines, we consider a linear regression model (LR) which predicts each output independently and an ARMAX model with AR(3) and MA(2) components, both using the same input features as the SGCRF.

A. Probabilistic predictions

Typically, forecasting systems are evaluated solely on the quality of the point forecasts produced and we see in Table I that, on this basis, the SGCRF method performs significantly better than linear regression and ARMAX. Due to the high-dimensionality of the feature space relative to the number of training examples, we expect the ℓ_1 penalty employed by the SGCRF to be statistically efficient in identifying the underlying structure of the correlations in wind power production and the dependence of wind power on wind forecasts; indeed in Figure 1, we see that the estimated parameters exhibit a high degree of sparsity.

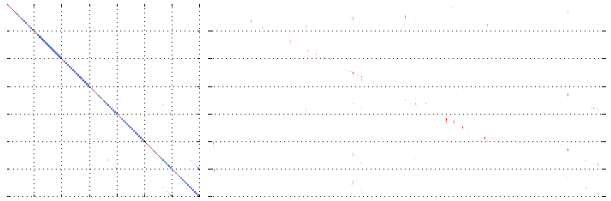

Fig. 1. Sparsity patterns Λ and Θ from the SGCRF model. Λ is estimated to have 1412 nonzero entries (1.2% sparse) and Θ is estimated to have 7714 nonzero entries (0.67% sparse). White denotes zero values and wind farms are grouped together in blocks.

TABLE II

COVERAGE OF CONFIDENCE INTERVALS

Method	Task	90%	95%	99%
LR	Aggregate farms	0.6943	0.7653	0.8600
	Aggregate times	0.3790	0.4451	0.5364
	Both	0.1944	0.2500	0.3333
LR + copula	Aggregate farms	0.7256	0.8051	0.9040
	Aggregate times	0.4028	0.4663	0.5728
	Both	0.2176	0.2639	0.3380
ARMAX	Aggregate farms	0.5682	0.6473	0.7570
	Aggregate times	0.6779	0.8188	0.9544
	Both	0.2454	0.3102	0.4213
SGCRF	Aggregate farms	0.8267	0.8791	0.9443
	Aggregate times	0.6104	0.6885	0.7976
	Both	0.4306	0.5370	0.6389
SGCRF + copula	Aggregate farms	**0.8981**	**0.9468**	**0.9830**
	Aggregate times	**0.8743**	**0.9266**	**0.9722**
	Both	**0.8796**	**0.9259**	**0.9676**

However, we are primarily interested not in the accuracy of point forecasts, but in the ability of the models to capture the distribution of future power production. Indeed, as shown in the same Table I, the inclusion of the copula transform degrades the performance of the models in terms of RMSE; this is expected since by assuming a Gaussian distribution over the noise, the untransformed models are explicitly minimizing mean squared error. RMSE alone is a poor measure of how well an algorithm can actually predict future observations: if we judge the algorithms by the ability to accurately predict the range of possibilities for future outcomes, a different picture emerges. For example, a natural task for a wind farm operation would be to generate a distribution over *total* power produced by all seven wind farms in the next 24 hours, in order to establish 95% confidence intervals about the power to be produced; this could in turn by used by stochastic optimal dispatch method, to determine how much power to generate from other sources.

Table II illustrates the *coverage* of the confidence intervals generated by different approaches, evaluated on a held-out test set of the wind power data. For each example in the test set, we used each method to generate many samples of upcoming wind power and for each of these samples, we computed the total power aggregated over all the farms, all times, or both, and used these to generate histograms of the aggregated power. Finally, we used these histograms

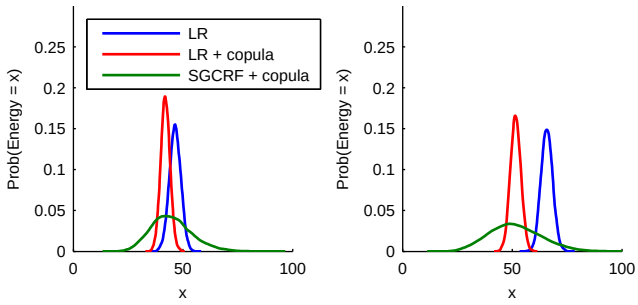

Fig. 2. Examples of predictive distributions for total energy output from all wind farms over a single day.

to estimate 90%, 95%, and 99% confidence intervals of the aggregate power, and evaluated how often the true total wind power fell into that interval.

As seen in Table II, the SGCRF + copula model produces intervals that map very closely to their desired coverage level, whereas linear Gaussian and ARMAX models perform much worse. To see why this occurs, we show in Figure 2 several of these estimated distributions of aggregate power, sampled from the different models. The independent Gaussian models are substantially overconfident in their predictions, as multiple i.i.d. random variables will tend to tighten the variance, leading to vastly inaccurate predictions when those variables are in fact highly correlated. We note that we could also consider a joint linear model by forming the unregularized MLE estimate for the covariance matrix, but in general this is not well-suited for high-dimensional problems and in fact is undefined for $p > m$. The ARMAX model does capture some of the correlation across time via the moving average component and we see in Table II that it performs significantly better than linear regression when aggregating predictions across multiple times. However, the SGCRF with the copula transform clearly achieves the best results implying that it is more accurately capturing the disperse nature of the actual joint distribution.

B. Application to ramp detection

One particularly appealing possibility for the probabilistic forecasting methods is in the area of predicting wind "ramps," times when power experiences a sudden jump from a relatively low to a relatively high value. Because wind power grows with the cube of wind speed in Region 2 of the turbine operating conditions (before the wind turbines reach rated power), a small increase in wind speed can lead to a large change in power, and predicting when these ramps occur is one of the primary open challenges in wind forecasting. Indeed, a well-known issue with many forecasting methods is that while they may accurately predict that a ramp will occur, they are significantly limited in accurately capturing the uncertainty over where the ramp will occur [4].

Although a detailed analysis of the ramp prediction capabilities of our approach is not the main focus on this paper, we briefly highlight the potential of our approach in this task. In particular, because the model accurately

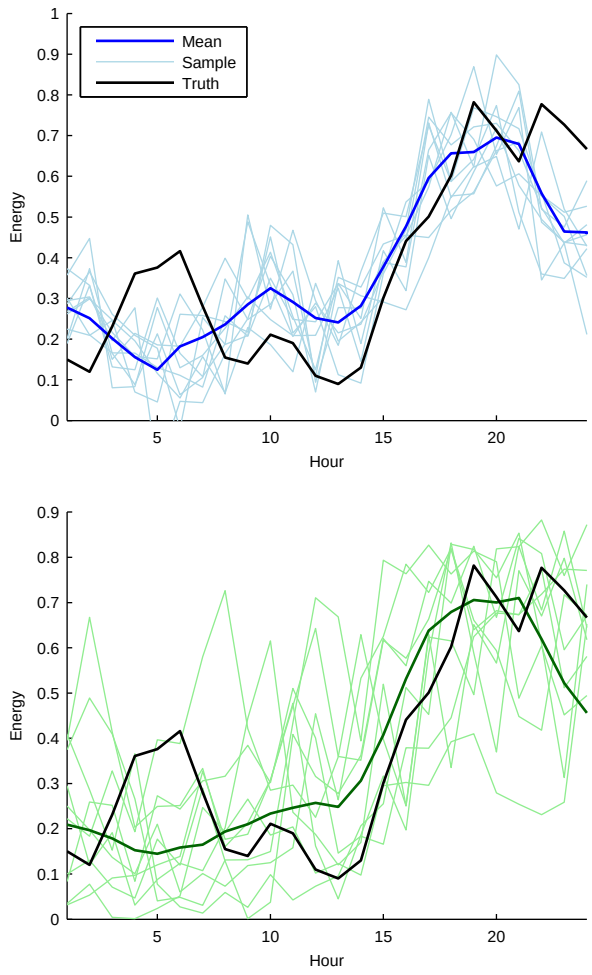

Fig. 3. Samples drawn from the linear regression model (top), and SGCRF model (bottom)

captures correlations in the predicted observations over time, if we draw random samples from our model, then we expect scenarios where the ramp occurs at different times; this is in contrast to most "independent" probabilistic methods, which would assume a fairly tight distribution over possible future scenarios. This situation is illustrated in Figure 3, where we show the mean prediction along with 10 samples drawn from each model. Again, in a stochastic optimal control task, an operator would be much better served by considering the possible scenarios generated from our model than from the independent probabilistic model, as they consist of several different timings for the upcoming power ramp.

V. Conclusion

In this paper, we present a probabilistic forecasting approach based on state-of-the-art methods in machine learning that can efficiently model high-dimensional and non-Gaussian joint distributions over its predictions. Such probabilistic forecasting is capable of more accurately capturing the true distribution and can be an indispensable tool for making accurate predictions of the actual range of possible observations. From a larger perspective, these methods highlight a highly desirable direction for future work: new machine learning approaches that can produce high-dimensional probabilistic forecasts. But to fully exploit these models, we also need stochastic dispatch and optimal power flow methods than can exploit this uncertainty to mitigate risks inherent with renewable intermittency. Integrating these planning approaches with probabilistic models thus seems to be a key direction for future work in power systems operation.

References

[1] N. Amjady, F. Keynia, and H. Zareipour. Short-term load forecast of microgrids by a new bilevel prediction strategy. *Smart Grid, IEEE Transactions on*, 1(3):286–294, 2010.

[2] 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*, 9:485–516, 2008.

[3] P. Brockwell. *Time Series Analysis*. Wiley Online Library, 2005.

[4] C. Ferreira, J. Gama, L. Matias, A. Botterud, and J. Wang. A survey on wind power ramp forecasting. Technical report, Argonne National Laboratory (ANL), 2011.

[5] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[6] T. Hong. Global energy forecasting competition, 2012.

[7] B. Lange, K. Rohrig, F. Schlögl, Ü. Cali, and R. Jursa. Wind power forecasting. *Renewable electricity and the grid*, pages 95–120, 2008.

[8] H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10:2295–2328, 2009.

[9] J. Mendes, R. Bessa, H. Keko, J. Sumaili, V. Miranda, C. Ferreira, J. Gama, A. Botterud, Z. Zhou, and J. Wang. Development and testing of improved statistical wind power forecasting methods. Technical report, Argonne National Laboratory (ANL), 2011.

[10] M. Milligan, M. Schwartz, and Y. Wan. Statistical wind power forecasting models: results for us wind farms. Technical report, National Renewable Energy Laboratory, Golden, CO, 2003.

[11] C. Monteiro, R. Bessa, V. Miranda, A. Botterud, J. Wang, G. Conzelmann, et al. Wind power forecasting: state-of-the-art 2009. Technical report, Argonne National Laboratory (ANL), 2009.

[12] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. 2002.

[13] J. Penm, J. Penm, and R. Terrell. The recursive fitting of subset varx models. *Journal of Time Series Analysis*, 14(6):603–619, 2008.

[14] P. Pinson. Estimation of the uncertainty in wind power forecasting. *Centre Energétique et Procédés–Ecole des Mines de Paris Rue*, 23, 2006.

[15] G. Sideratos and N. Hatziargyriou. An advanced statistical method for wind power forecasting. *Power Systems, IEEE Transactions on*, 22(1):258 –265, feb. 2007.

[16] K.-A. Sohn and S. Kim. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *Proceedings of the Conference on Artificial Intelligence and Statistics*, 2012.

[17] S. A. Soliman and A. M. Al-Kandari. *Electrical Load Forecasting: Modeling and Model Construction*. Elsevier, 2010.

[18] C. Sutton and A. McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2012.

[19] Various. *PJM Manual 19: Load Forecasting and Analysis*. PJM, 2012. Available at: http://www.pjm.com/planning/resource-adequacy-planning/ /media/documents/manuals/m19.ashx.

[20] M. Wytock and J. Z. Kolter. Sparse Gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In *Proceedings of the International Conference on Machine Learning*, 2013.

[21] X.-T. Yuan and T. Zhang. Partial gaussian graphical model estimation. *CoRR*, abs/1209.6419, 2012.

[22] I. Žežula. On multivariate gaussian copulas. *Journal of Statistical Planning and Inference*, 139(11):3942–3946, 2009.